THE MATHEMATICAL ASSOCIATION OF AMERICA

CONICS

THE DOLCIANI MATHEMATICAL EXPOSITIONS #29

Keith Kendig

Conics

© 2005 by
The Mathematical Association of America (Incorporated)
Library of Congress Catalog Card Number 2005923496

ISBN 0883853353

Printed in the United States of America

Current printing (last digit):
10 9 8 7 6 5 4 3 2 1

The Dolciani Mathematical Expositions

NUMBER TWENTY-NINE

Conics

Keith Kendig

Published and distributed by
The Mathematical Association of America

DOLCIANI MATHEMATICAL EXPOSITIONS

Council on Publications
ROGER NELSEN, *Chair*

Dolciani Mathematical Expositions Editorial Board
UNDERWOOD DUDLEY, *Editor*
ROBERT DEVANEY
JERROLD GROSSMAN
BENJAMIN KLEIN
VIRGINIA KNIGHT
MARK PETERSON
JAMES TANTON
WILLIAM S. ZWICKER

The DOLCIANI MATHEMATICAL EXPOSITIONS series of the Mathematical Association of America was established through a generous gift to the Association from Mary P. Dolciani, Professor of Mathematics at Hunter College of the City University of New York. In making the gift, Professor Dolciani, herself an exceptionally talented and successful expositor of mathematics, had the purpose of furthering the ideal of excellence in mathematical exposition.

The Association, for its part, was delighted to accept the gracious gesture initiating the revolving fund for this series from one who has served the Association with distinction, both as a member of the Committee on Publications and as a member of the Board of Governors. It was with genuine pleasure that the Board chose to name the series in her honor.

The books in the series are selected for their lucid expository style and stimulating mathematical content. Typically, they contain an ample supply of exercises, many with accompanying solutions. They are intended to be sufficiently elementary for the undergraduate and even the mathematically inclined high-school student to understand and enjoy, but also to be interesting and sometimes challenging to the more advanced mathematician.

---

1. *Mathematical Gems,* Ross Honsberger
2. *Mathematical Gems II,* Ross Honsberger
3. *Mathematical Morsels,* Ross Honsberger
4. *Mathematical Plums,* Ross Honsberger (ed.)
5. *Great Moments in Mathematics (Before 1650),* Howard Eves
6. *Maxima and Minima without Calculus,* Ivan Niven
7. *Great Moments in Mathematics (After 1650),* Howard Eves
8. *Map Coloring, Polyhedra, and the Four-Color Problem,* David Barnette
9. *Mathematical Gems III,* Ross Honsberger
10. *More Mathematical Morsels,* Ross Honsberger
11. *Old and New Unsolved Problems in Plane Geometry and Number Theory,* Victor Klee and Stan Wagon
12. *Problems for Mathematicians, Young and Old,* Paul R. Halmos
13. *Excursions in Calculus: An Interplay of the Continuous and the Discrete,* Robert M. Young
14. *The Wohascum County Problem Book,* George T. Gilbert, Mark Krusemeyer, and Loren C. Larson
15. *Lion Hunting and Other Mathematical Pursuits: A Collection of Mathematics, Verse, and Stories by Ralph P. Boas, Jr.,* edited by Gerald L. Alexanderson and Dale H. Mugler
16. *Linear Algebra Problem Book,* Paul R. Halmos
17. *From Erdős to Kiev: Problems of Olympiad Caliber,* Ross Honsberger
18. *Which Way Did the Bicycle Go? ...and Other Intriguing Mathematical Mysteries,* Joseph D. E. Konhauser, Dan Velleman, and Stan Wagon
19. *In Pólya's Footsteps: Miscellaneous Problems and Essays,* Ross Honsberger
20. *Diophantus and Diophantine Equations,* I. G. Bashmakova (Updated by Joseph Silverman and translated by Abe Shenitzer)
21. *Logic as Algebra,* Paul Halmos and Steven Givant
22. *Euler: The Master of Us All,* William Dunham
23. *The Beginnings and Evolution of Algebra,* I. G. Bashmakova and G. S. Smirnova (Trans. by Abe Shenitzer)

24. *Mathematical Chestnuts from Around the World*, Ross Honsberger
25. *Counting on Frameworks: Mathematics to Aid the Design of Rigid Structures*, Jack E. Graver
26. *Mathematical Diamonds*, Ross Honsberger
27. *Proofs that Really Count: The Art of Combinatorial Proof*, Arthur T. Benjamin and Jennifer J. Quinn
28. *Mathematical Delights*, Ross Honsberger
29. *Conics*, Keith Kendig

MAA Service Center
P. O. Box 91112
Washington, DC 20090-1112
1-800-331-1MAA fax: 301-206-9789

In memory of
Hassler Whitney

Photo: Herman Landshoff. Courtesy of the Archives of the Institute for Advanced Study.

Hassler Whitney
1907–1989

Preface

Beauty is only skin deep, but
Ugliness goes all the way
to the bone.
—Anonymous

In mathematics,
Ugliness is only skin deep, but
Beauty goes all the way
to the bone.
***—Philosopher,* in this book**

This book arose from a simple philosophy: in mathematics, the fundamental ought to be beautiful. Conics are certainly basic, arising in many areas of mathematics as well as in physics. Their theory should be consistent, unified and beautiful. Throughout years of teaching conics at various levels—from pre calculus to graduate engineering—I all too often bumped into a lack of consistency, unity and beauty. I found that much of the traditional treatment is more of a collage of smaller designs rather than a unified whole. For example, formulas and routines used for ellipses are different for hyperbolas, and different yet for parabolas.

There also seem to be gaps in the general body of knowledge about conics. As just one example, after we talked in class about tilted conics $Ax^2+Bxy+Cy^2=1$, a student asked if there is a formula giving the area of a tilted ellipse. What a *natural*! The coefficients completely determine the tilted conic, so why not a formula in terms of A, B and C? I'd never seen such a formula, so I ended up figuring it out: Area $= 2\pi i/\sqrt{B^2-4AC}$. I eventually noticed that if the tilted conic is not an ellipse but a hyperbola, plugging in the coefficients A, B and C still produces an answer. Now, what on earth did "the area of a hyperbola" mean? The formula was supplying information about *something*. What was that something?

When it was suggested that I write a book for the MAA, I saw an opportunity to finally "do conics right." Perhaps I could at last see conics in a more unified way, and fill in some of those gaps. This book is my answer.

The book is written as a "mini-seminar" involving three characters. To an extent their question-and-answer process mirrors my own way of working—lots of examples, all sorts of little debates on why things don't seem beautiful, and what to do about that state of affairs.

Our characters have varying amounts of mathematical background and quite different personalities, but this lends a synergism to their explorations. Here they are:

Philosopher is by far the most iconoclastic of the three, and progress couldn't take place without him. He's the main source of important questions, and these generally come to him because of his unusual radar system. Like a good detective, he picks up subtle clues often leading to larger consequences. Think, for example, of a paleontologist working at a site, whose experienced eye picks up a small outcropping that turns out to be a part of a buried dinosaur.

Philosopher's "outcroppings" are things like arbitrariness, lack of symmetry and exceptions, all frequently lying buried in centuries-old nomenclature, specific-use formulas or points of view born of an age when understanding was more limited and expectations different.

Teacher is mathematically the most broadly educated of the three, but tends toward the traditional. But with his extensive background, he is often in a position to make important connections, and can see a broad landscape developing. He has an open mind and is quick to see value in another's point of view. He is, in fact, often inspired by Philosopher's sometimes outrageous questions. And when an idea requires a fuller explanation or clarification, he is only too happy to oblige.

Student is, as his name implies, the least experienced of the trio. But his fresh spirit and uncluttered mind often prove very useful. He appreciates the power of concrete examples and often asks for them; these can lead to important insights. When the others seem to get just a bit too highfalutin or abstract, it is he who brings them back to *terra firma*.

Our trio attempts to rectify their problems one by one, and their journeys lead them to an increasingly beautiful and expansive vista of conics.

What are the prerequisites? This book is written mainly for upper-division students and beyond, although less experienced students will learn something new. It could form the basis of a senior capstone course. The book assumes a course in calculus, basic linear algebra, some differential equations and at least a popular exposure to "coffee-cup-and-donut topology." For those who don't know what a topological space is, Appendix 2 fills in the basics. Also, there are a few exercises requiring a rudimentary ability to use Maple, Mathematica or other software of this type.

I've attempted to write the book so that if you don't understand something, and if the above appendices don't help, then it's okay just to skip it—that won't stand in the way of getting the main idea.

And those boxes? There are shaded boxes scattered throughout the text. These contain anything from historical tidbits, to comments illuminating or synopsizing what we've been talking about, to the occasional pearl of wisdom. Mostly, those bits of wisdom are intended for everyone, but occasionally one is aimed at the graduate student or beyond.

Acknowledgments. I want to express my appreciation to a number of people who helped me with this book, directly or indirectly. My wife Joan has been a constant and helpful companion during the nearly two years I've spent writing this book. She read the entire manuscript, making many suggestions for improvement, thanks to her English-major and book-writing background. I so much appreciate everything! And my son, Chris, gave me a tremendous amount of help: he put together a web site for me, helped me learn about and make applets, and supplied 24/7 tech support. All that followed our building together a super-fast, capacious and smoothly-working computer which eventually earned the house-

hold nickname "Honcho." Out of sight and sound in the basement, with connections to the first floor, it has quietly made the many technical aspects of writing a book almost fun.

Many, many thanks to Don Albers who approached me about writing a book after the Ford Prize presentation. I was sort of in another world, and he must have sensed it. But he kept at it until I finally woke up and said to him, "You're *serious*, aren't you?" The result is this book. Throughout the entire project, he has been a constant and upbeat "cheerleader"—a real friend.

I also owe a big debt of gratitude to Dan Velleman, the outstanding editor of the Dolciani Series. He read carefully for content and understanding, uncovering intellectual glitches, including not only things I said wrong, but things I didn't say but should have. His comments were to-the-point and insightful. I was fortunate to have his great care go into this book.

I knew there would be many figures in my nascent *opus*, and I wanted to do things right: there *had* to be some "best way" to create all those figures. In writing short papers, I'd muse from time to time about that problem, but had never encountered the right person to ask—that is, until I happened to mention it to Beverly Ruedi, manager of electronic production at MAA. I'll never forget the experience. After that very enlightening phone conversation, Joan and I dashed down to the university bookstore, arriving about ten minutes before closing, and got a copy of Adobe Illustrator 10. I soon began to sense the absolutely incredible power of this program working in tandem with software such as Mathematica, which creates mathematically correct vector-based graphics. Such output seamlessly exports to Illustrator for polishing in almost any direction. Of course, there was a learning curve for me, but as time went on, I felt my "illustrating powers" grow naturally. For me, it's been a good experience, and I hope you like the result.

It is also a pleasure to say "Thank you!" to Elaine Pedreira, who efficiently and skillfully performs the multitude of tasks in shepherding MAA manuscripts through production. She edits copy for ads, contracts for compositors, artists and printers, and solves problems—big and little—that always have a way of arising unbidden. It has been my good fortune to have benefited from her long and broad experience.

My colleagues at Cleveland State have been very supportive. John Oprea fielded what must have seemed like an endless stream of questions from me as I struggled to fill in yawning chasms in my knowledge of differential geometry. He carefully read Chapters 13 and 14 and made many helpful suggestions. Barbara Margolius read other parts of the manuscript and gave me the main impetus to include applets. That meant learning something new, and I want to thank both Sally Shao and Felipe Martins for patiently helping me through some tricky aspects in making applets work.

In connection to applets, "JavaView" is a marvelous graphics add-on to Maple or Mathematica, developed by Steve Durango, Konrad Polthier and others in Germany. Professor Polthier answered my many e-mail inquiries promptly and informatively, and it is a real pleasure to say "Thank you!" for his tremendous support.

Some of my former math teachers and colleagues have left an indelible mark. It was my high school teacher Robert Crawford who inspired me to go on in mathematics. I'll never forget the excitement I felt the time he explained how dragging an ellipse up to the line at infinity makes a parabola, and crossing that line makes a hyperbola. Wow! His enthusiasm and love of mathematics were contagious. He was blunt, outspoken and tremendously supportive. As an original, mold-busting character, it's not entirely outrageous to sum up Bob Crawford this way: Feynman was the Feynman at CalTech; Crawford was the Feynman at Santa Monica High School.

Angus Taylor was my calculus teacher at UCLA; watching his simple, personable approach to solving calculus problems made a lifelong impression. Once after class, when I wasn't too clear on some point, he asked me to explain my understanding of it as best I could. I struggled, he patiently listened, and after a while the problem began to solve itself. Then he asked me if I had a younger brother. No, I didn't. "Well, make one up, and explain things to him, in his terms, so he finally understands!" I never forgot my teacher's advice, and some readers will pick up his influence in this book.

Basil Gordon was my thesis advisor at UCLA. I first met him in a class on tensor methods given by Rolf Nevanlinna, and it was soon clear that this young faculty member was not only uncommonly bright, but had a special talent for clearly explaining things when I got confused. It turned out that the next year he was to teach a course on algebraic geometry, and I jumped at the chance to enroll in it. He turned me on with his very first lecture—drag two crossing ellipses, and they still intersect in four points—it's just that the points go off into complex space of four real dimensions. If we could see in 4-D, we could follow them along. His many subsequent models-of-clarity lectures made my going into algebraic geometry all but inevitable.

Basil has remained a true friend and supporter throughout the many years. He not only proofread *Elementary Algebraic Geometry*, but this book as well. I was so happy to benefit once again from a rare combination of a broad and deep mathematical background, sharp editorial eye, and a wonderful sense of language. His careful reading has resulted in many dozens of corrections and improvements, and his unstinting care has indeed left a strong imprint on *Conics*.

I worked with Hassler Whitney for two years at the Institute for Advanced Study, and the mark he left on me goes deep. I guess I'm attracted to unconventional, original sorts, and Hass was surely that. His undergraduate days were at Yale, so one might expect that here was just about the most incredible math major ever. But his major was music, not math. Well, then, he must have taken all sorts of math courses, and.... Actually, he took almost no math courses. Mathematically, he was largely self-taught.

Anyway, one day, in his office, I happened to mention Bézout's theorem, which basically says that two curves of degree n and m intersect in nm points. He says he never heard of it (Bézout's theorem is in fact highly under-appreciated), and seems galvanized by it. He jumps up and heads toward the blackboard, saying "Let's see if I can disprove that!" *Disprove it?!* "Wait a minute!" I say, "That theorem is nearly two centuries old! You can't disprove anything...really..." As he begins working on some counterexamples at the blackboard, I see that my well-meant words are simply static.

His first tries were easy to demolish, but he was a fast learner, and ideas soon surfaced about the complex line at infinity, and how to count multiple points of intersection. After a while, it got harder for me to justify the theorem, and when he asked, "What about two concentric circles?" I had no answer. He argued his way through, and eventually found all four points. Finally he was satisfied, and the piece of chalk was given a rest. He backed away from the blackboard and said, "Well, well—that *is* quite a theorem, isn't it?"

I think I mostly kept my cool during all this, but after I left his office, I realized I was pretty shaken. I remember thinking to myself, "Golly, Kendig, you just saw how one of the giants does it!" He'd taken the theorem to the mat, wrestled it, and the theorem won. I'd known about that result for at least two years, and I realized that in 15 or 20 minutes, he'd gained a deeper appreciation of it than I'd ever had. In retrospect, it represented a turning point for me: I began to think examples, examples. Whitney worked by finding an example that contained the essential crux of a problem, and then worked relentlessly on it until he cracked it. He left it to others to generalize. It is to Hass that I affectionately dedicate this book.

Contents

with brief chapter descriptions

1 Beauty and the Beast **1**

Philosopher, attempting to learn about conics, quickly becomes frustrated by the lack of symmetry and many exceptions encountered in the standard treatments. He pours his heart out to Teacher, whose patience and broad knowledge help the trio arrive at the unification and greater depth inherent in the subject. This is usually done by insisting that things be consistent, symmetric and beautiful.

2 Life at Infinity **27**

It is discovered that an important source of problems lies in the traditional "viewing screen" of conics, the plane. By drawing identical ellipses on opposite sides on a sphere and then rotating the sphere, one sees that as part of one ellipse disappears over the horizon, another part of its mate appears over the opposite horizon, thus creating the two branches of a hyperbola. The horizon, in mathematical terms, turns out to be the "line at infinity."

3 How to Gift-Wrap a Hyperbola **45**

Every ellipse symmetrically rests inside its "standard rectangle," touching the midpoint of each of the rectangle's four sides. If ellipses and hyperbolas are so unified, then why don't we see a similar box surrounding a hyperbola? Standard treatments never answer this. By looking at conics in projective space, it becomes apparent that the familiar diagonals, used in drawing a hyperbola's asymptotes, are in fact the two missing rectangle sides.

4 The Cube **57**

The sphere is compressed to form a cube, and the ellipse together with the two standard hyperbola views (corresponding, roughly, to equations $x^2 + y^2 = 1$, $x^2 - y^2 = 1$ and $-x^2 + y^2 = 1$) now appear as circles on the three pairs of opposite faces. This not only highlights inherent symmetry, but the flat faces of the cube make viewing curves drawn on them easy and familiar.

5 The Other Foci: A Well-Kept Secret **65**

A ideally-reflecting ellipse has two "sweet points," where light or sound emanating from one focal point converges to the other. Our trio of characters makes its first really major

discovery: by expanding the context from the real domain to the domain of complex numbers, they discover a new pair of focal points lying on the imaginary axis. All the traditional properties of focal-point pairs are shown to hold for this new pair.

6 Are Hyperbolas *Really* Ellipses? 89

By this time, the trio has fairly well established that hyperbolas and ellipses are one and the same—just looked at from different perspectives. Our friends now perform a series of reality checks, showing that computations carried out for an ellipse directly translate to "mirror" computations for a hyperbola, and vice-versa.

7 Stakes and Strings 103

Tie each end of a piece of string to a thumb tack and push the tacks into a drawing board, leaving some slack in the string. Then use a pen to keep the string taut and swing the pen around. This draws an ellipse, and since the string doesn't stretch, the two string segments always sum to the same constant. To get a hyperbola, you subtract the two string lengths. But if hyperbolas are essentially ellipses, then why must you subtract? The trio discovers that graphing the distance from either focus to an ellipse point produces a line. This leads to signed distances, and not only solves the ellipse-hyperbola conundrum, but leads to an exciting discovery: as you move beyond an ellipse vertex, one segment keeps on giving to the other, and you climb up a hyperbola branch in the Minkowski plane. When one string has donated all its real length to the other, the second string, having zero length, lies on a null line.

8 Directrices, New and Old 121

One way the ancient Greeks defined conics was via a fixed point (focus) and line (directrix). Philosopher's aesthetic sense tells him that since there's a directrix associated with each focus, and since they'd found a new pair of foci, there should also exist a second pair of directrices. By carefully constructing drawings of 3-D slices of four-space, they succeed in uncovering the suspected new pair of directrices, and they show that these directrices have properties analogous to traditional ones. This removes some further exceptions that had been bothering Philosopher.

9 Conics in General Position 143

Any ellipse can be looked at as a suitably-stretched circle. Is the same true of hyperbolas? The trio uses the power of linear algebra to show that it is. They come to realize that symmetric matrices can have imaginary entries, and find that it's precisely stretching in imaginary directions that produces hyperbolas.

10 A Beautiful Mathematical Universe 169

Teacher explains how to add points at infinity to $\mathbb{C}^2$, arriving at "complex projective 2-space," $\mathbb{P}^2(\mathbb{C})$. This turns out to be the best setting in which to consider conics. In a real, finite setting, intersection points can seem to vanish. For example, take two "crossing ellipses," meeting in four points. As you move one ellipse but not the other, those four points decrease to two after a while, and ultimately, to no points of intersection. The trio discovers that the "disappearing points" never actually disappear at all—they simply move

into complex space, sometimes even escaping to infinity. The larger setting keeps track of all points, and this leads to simple, consistent results.

11 A Most Excellent Theorem **183**

In moving one of the above "crossed ellipses" precisely when the number of points decreases, the ellipses become tangent to each other, two points coalescing into one. By regarding such merged points as "multiple"—lying on top of each other—one can prove a very general and beautiful extension of the fundamental theorem of algebra. The trio explores some of the consequences of this generalization, including Pascal's and Brianchon's theorems.

12 The Big View **215**

At this stage, the trio has extended its vistas to the larger "living space" of $\mathbb{P}^2(\mathbb{C})$. Since there are now more solutions to the equations defining conics, the conics not only include more points, they also now live in four dimensions. Can we peek into the fourth dimension to see what these more complete conics look like? By gluing together 3-D slices of 4-space, the trio finds that these conics are topologically spheres.

13 Curvature **235**

To gain a better understanding of just how conics are situated in four dimensions, the trio works out curvature formulas for general smooth surfaces.

14 Curvature of Conics **269**

The formulas obtained in Chapter 13 are now applied to conics themselves. The trio discovers that the usual circle, as it lies on the full conic—a topological sphere in 4-space—does indeed have constant curvature. But that curvature is negative, not positive. Locally, the circle looks like the innermost circle drawn on an inner tube.

15 Photons and Conics **283**

One problem has bothered Philosopher ever since he first began studying conics: if you release a photon from one focus of a reflecting ellipse, then when it hits the ellipse, it bounces off and heads directly to the other focus; but try that with a hyperbola, and the photon, after reflecting, always *heads the wrong way, toward "outer space"!* Textbooks never address this issue, but it's a problem Philosopher finds he simply must resolve. The trio discovers that it is once again the plane that causes the difficulty. Replace the plane with a very large sphere, and let the photon do the physically correct thing. The photon will then travel along a long geodesic (a "great circle"), finally arriving at the other focus, as it should.

16 How Conics Solved a 2000-Year-Old Problem **305**

One of the most torturous scientific odysseys was the centuries-long journey to unravel the mystery of planetary motion. Kepler, of course, discovered that their orbits are ellipses. Newton found something that Kepler missed: heavenly bodies may also move along parabolas or hyperbolas—conic sections. His masterpiece *Principia* included a way to find the equation for Kepler's ellipse from a bare minimum of observational data. After review-

ing Newton's arguments, the chapter gives the explicit equation of an elliptical orbit in terms of the "launch data."

17 Waves and Conics **341**

Take an electric tuning fork having a needle soldered to the end of each prong, and dip the vibrating needle points into a vat of quiet water. The circular ripples created by each vibrating needle interfere, forming a system of hyperbolic "interference fringes" that are broader than the circular background ripples. In the practical world, light from a star passing through two slits in a baffle creates a series of light-and-dark fringes, and the apparatus can be arranged to make these fringes far larger than the wavelength of light. This is the basis of modern radio astronomy, which has provided pictures of unprecedented detail of galaxies near the boundary of the universe, of quasars, and of details near the black hole at the center of our Milky Way. The trio explores these timely and exciting ideas.

Appendix 1 Some Conics Formulas **359**

A general conic has five degrees of freedom (three if its location is fixed—for example, if its center is at the origin). Since these parameters uniquely determine the conic, they uniquely determine measurements like eccentricity, principal-axis lengths, latus rectum, area, distance from center to a focus, distance from center to a directrix, and principal curvatures. We give formulas for these measurements for eight different ways of specifying a conic: intersection of a cone and plane; polar equation; focus-directrix-eccentricity; stake-and-string construction; parametric equations of a conic; conic defined by a quadratic form; path of a celestial body; conic through five points.

Appendix 2 Topology: A Quick Handshake **379**

Elementary topological notions appear at several places in the text. This appendix gives the necessary topological definitions, with examples, for those who need them.

Suggestions for Further Reading **393**

Applets **397**

Index **401**

About the Author **403**

1

Beauty and the Beast

in which departures from traditional thinking help expose some unsuspected beauty

Beauty vs. Reality

Teacher. I hear you've been studying conic sections. Conics go back more than 2300 years, and have found their way into almost every area of mathematics and physics. What a beautiful subject!

Philosopher. Beauty was my vision. Then came the reality.

Teacher. What do you mean?

Philosopher. My journey into conics started very nicely. But then I got to thinking about how the whole subject ought to be perfectly unified. It was then that I began running into a number of simple-sounding questions. They all ought to have simple and obvious answers—I just can't answer them!

Teacher. What got you to thinking that the whole subject should be perfectly unified?

Philosopher. It happened when I saw that among nondegenerate conics, there's "conservation of information," implying that they're all essentially equivalent.

Teacher. I don't understand.

Philosopher. Well, every conic is the intersection of some double cone and a plane. If the plane goes through the cone's vertex, then the conic is *degenerate*—like a line, two lines or a point, for example. It's when the plane misses the vertex that you get nondegenerate conics—true ellipses, parabolas or hyperbolas.

Teacher. I agree with that.

> Euclid defined conics using only a single cone. It was Apollonius (ca. 262–ca. 190 BCE) who introduced the double cone, giving the hyperbola its two branches.

Philosopher. If you keep track of where the vertex was, then for nondegenerate conics you can go back and reconstruct the original cone—just pass lines through the vertex and the points of the conic. You might need to take the topological closure of this, but the essential thing is that you can always retrieve the cone. (Topological closure is defined in Appendix 2.)

Cone comes from the Greek *konos*, meaning "pine cone"; typically, an end of a pine cone has a conical shape. *Konos* is thought to be also related to our word "hone." A pencil that is honed to a point has a conical shape—think of a pencil sharpener.

Apollonius
ca. 262–ca. 190 BCE

Apollonius of Perga wrote an extraordinary treatise, *Conic Sections*. This eight-book masterpiece completely superseded works on conics by earlier geometers such as Euclid, Aristaeus and Menaechmus. It was Apollonius who introduced today's method of defining a conic by slicing a double cone with a plane. Before him, geometers used three separate "flavors" of single-napped cone, having vertex angle less than, equal to or greater than a right angle; the slicing plane was then always perpendicular to a line in the cone. The respective section was then an ellipse, parabola or one branch of a hyperbola. Apollonius realized that a hyperbola needed to have two branches, and his cone sections naturally produce them. In many ways Apollonius' work represented, in its originality, depth and completeness, the pinnacle of ancient mathematics. It earned him the nickname "The Great Geometer."

Teacher. This doesn't work for degenerate ones?

Philosopher. No. For example, there are lots of different cones intersecting a plane in a given point. Likewise, many cones can intersect a plane in one line. Same thing for

two crossing lines. But nondegenerate conics? They're all blood brothers of each other—Figure 1.1 suggests the idea:

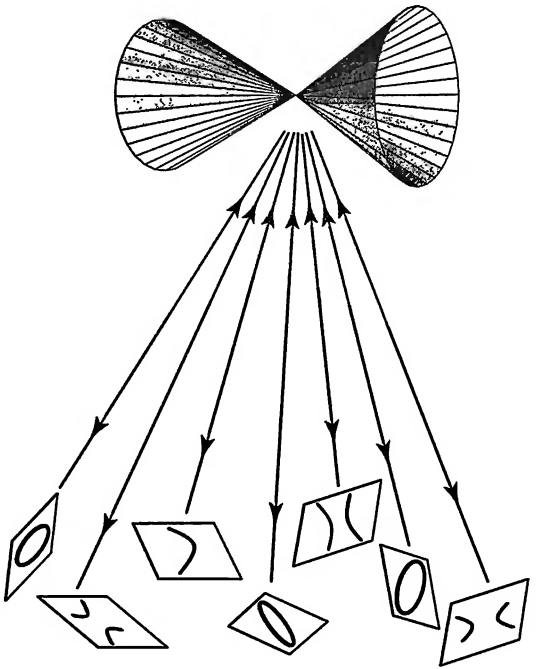


FIGURE 1.1. At the top is the parent cone and at the bottom, a sampling of nondegenerate conics. Any down arrow represents taking a nondegenerate slice of the cone. The up arrow from any one of the slices represents reconstructing the parent cone. All these nondegenerate conics are equivalent in the sense that any two of them are connected by a two-way path passing through the parent cone.

At the top we see the parent cone and below that, a sampling of ellipses, parabolas and hyperbolas that are plane sections of the cone. One could call them "children." Any downward arrow represents taking a section; any upward arrow represents reconstructing the cone from that section. The plane cutting the cone becomes a "screen," giving a particular view of the conic. Starting at the bottom of the figure, we can go up to the parent cone and then back down to another conic, and in this sense, the two conics are *equivalent*. From the standpoint of screens, they're simply different views of the same thing.

Teacher. This is a nice observation, but how does this lead to questions you can't answer?

Philosopher. Well, if all those nondegenerate conics are equivalent, then whatever you say about one should make sense with any other. So, for example, any ellipse encloses an area. But hyperbolas don't! One could say "Well, they just don't—that's obvious!" But I can't believe that this represents the ideal state of any theory. There's a big information gap.

Teacher. Your point of view could generate a lot of questions.

Philosopher. Tell me about it! As just one example, hyperbolas have asymptotes, so *why don't ellipses*? Again, the obvious answer: "That's a nonsensical question. Ellipses are bounded—there can't be asymptotes!" But if ellipses and hyperbolas are in fact different views of the same thing, and no information is lost, then where, exactly, do the asymptotes disappear to? Or how about this: when you start sketching an ellipse from its standard-form equation, it is natural to draw the "fundamental rectangle" that encloses it. I don't see a rectangle of *any* sort enclosing a hyperbola!

Teacher. I don't either, but maybe your questions aren't posed right.

Philosopher. I've wondered about that. But there are other, more specific and clear-cut questions, too. One of them, for example, comes from an ancient method of generating an ellipse—the "stakes and string" method: drive two stakes into the ground, and to each stake, tie one end of a piece of string. Now, keeping the string taut, trace out an ellipse. The stakes are the ellipse's foci, and the taut string segments can be thought of as the path of a photon as it goes from one focus to a point on the ellipse, reflects off the ellipse and proceeds to the other focus.

Student. This has a practical application: imagine an enormous ellipse, with the sun at one focus and the earth at the other. Here on earth, just a tiny part of this ellipse, maybe a foot or two wide, can intercept and reflect to the focal point enough radiant energy from the sun to start a fire!

> It was Johannes Kepler (1571–1630) who, in 1604, added *focus* to the vocabulary of conics. The term comes from the Latin for *hearth* or *fireplace*, which represents a room's focal point or center of interest.
>
> It does not refer to heat or burning.

Philosopher. But again, I run into a problem if I ask, "Why doesn't this work for a hyperbola?" If I release a photon from F_1, it bounces off the mirrored hyperbola and heads for "outer space." Forget about starting any fire!

> Conic sections were largely forgotten in the centuries following the ancient Greeks. It was only in Kepler's decades-long struggle to determine planetary paths that he turned—mostly in desperation—to those ancient curiosities. To him, the ellipse was a definite step down from his beloved "perfect" circles.

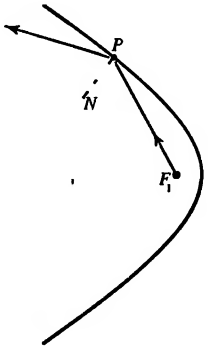


FIGURE 1.2. A photon released from focus F_1 reflects off a hyperbola the "wrong way"—it heads for the void rather than toward the other focus.

Student. But there is a standard picture for hyperbolas:

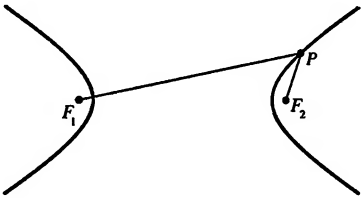


FIGURE 1.3. One can draw a "stakes and rubber band" model of a hyperbola. In it, $F_1P - PF_2 = K$, where the constant K is the shortest distance between the branches.

The segments seem to represent light paths for hyperbolas, too.

Philosopher. But you've got the photon piercing the left-branch mirror so it can get to P!

Student. We could throw that branch away, perhaps ...

Philosopher. That doesn't seem fair, but even if you did, then when the photon reaches P and reflects, it again heads off into outer space. To make things work, you've got to force the photon to go *backwards* to reach the other focus.

That is, to make the standard picture work, you have to break two physical laws: pierce one mirror, and make the photon move backwards as soon as it reflects off the other. That's crazy! So often, when the ellipse performs like a beauty, the hyperbola acts like a beast.

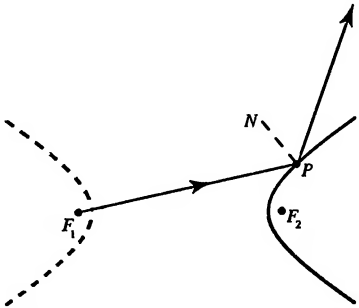


FIGURE 1.4. Even if the photon goes through the left branch, it still reflects off the right branch the wrong way. It would have to travel backwards to reach F_2.

Teacher. I admit there seems to be a problem here. Your contention is that because conics are in their innermost nature unified, what works for the goose, should work for the gander?

Philosopher. Yes. To me, the standard picture is wrong.

Teacher. But it's so well established!

Philosopher. It sure is. But I'm thinking that might be the fault of man, not of Mathematics itself. When I say this, I'm viewing things through "Platonic eyeglasses": Mathematics with a capital "M" is what's objectively out there, independent of anything we human beings might know of it. In contrast, mathematics with a lower-case "m" is what we happen to see or understand of it. We devise definitions, conventions, modes of representation and so on, in our struggle to understand Mathematics. The problem is that sometimes our constructs get set in concrete long before we arrive at a truer, more complete picture, and then those conventions work against us.

Teacher. I see your point. I guess that in some cases, those conventions could have worked against us, actually impeding the discovery process! When I was very young, I had an experience that sounds a lot like this.

Student. Do I hear a story coming? I love stories!

Teacher. It happened when I started learning about decimals in school. I knew then that ten has one zero, a hundred has two, a thousand, three, and so on. And then this teacher starts saying that a *tenth* doesn't have *any* zero, a *hundredth* has only *one*, a thousandth only two, and so on. I got very upset, and I told her she had to be wrong. She assured me that this is really the way things are. I felt betrayed by math. I loved its consistency, and the way you could figure things out, and this seemed like a darker side or something. Only much later did I have enough perspective to put my finger on the problem: **The decimal point is always misplaced!**

Philosopher. That's a pretty brave statement.

Teacher. Yes, but it means that Mathematics is still wonderful; mathematics may have betrayed me, but Mathematics didn't.

Student. You have to tell me how the decimal is misplaced!

Teacher. Well, the only essential role the decimal point plays is to identify the "units" position. For historical reasons—writing sequentially, printing or typesetting in a line—the decimal point got put to the right of the unit place. But it really shouldn't go there, any more than it should go to the left. One symmetric, fair place would be *over* the number. Then $1\hat{0}$ is ten and $\hat{0}1$ is a tenth. In fact it could go under, and it doesn't even have to be a point. It could be a bar, or the number in the units position could be written in bold, or color, or even with a circle around it—anything to make it special, but in a symmetric way:

$$1000^{*} \qquad {}^{*}0001$$

$$1000 \qquad 0001$$

$$100\boxed{0} \qquad \boxed{0}001$$

FIGURE 1.5. Any of these would be more symmetric than decimal notation for one thousand and one thousandth.

I get the feeling that, analogously, some of your conundrums may be coming from various long-established conventions and assumptions, just like the decimal point.

Philosopher. Wouldn't that be incredible? I'd be very obliged to begin exploring this approach with you. Maybe I can get unstuck! If we succeed, we'll have a more beautiful theory of conics.

> Compared to the decimal point, base ten is older and even more arbitrary. Arising from the peculiarity of ten digits (fingers), our base was set in stone long before Gottfried Leibnitz (1646–1716) discovered base two in 1679. Base two is less efficient than base ten but is mathematically more natural.

Viewing Screens

Teacher. Last time, we referred to the cutting plane as a "screen," and that got me thinking. It seems that the vertex of the double cone can be regarded as a light source, each line consisting of two oppositely-directed light rays. The plane is like the camera's film, and the conic is the resulting photograph.

Student. Sort of like shining a flashlight on a flat ceiling? You can get a circle, ellipse, or parabola.

Teacher. Yes, and you can get a hyperbola, too. Use a "double flashlight" like this:

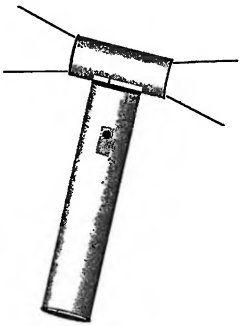


FIGURE 1.6. Two conical beams of light emanating from a "double flashlight."

Hold it so both beams hit the ceiling, and you'll get the two branches of a hyperbola.

There's a nice continuity too: if the cone of light is circular and the screen is perpendicular to the cone's axis of symmetry, the image is circular. But when you turn the flashlight a little bit away from perpendicularity, you introduce some distortion, and the image is elliptical. As you continue this you get more and more distortion, and turned far enough, a light ray finally becomes parallel to the ceiling and what you see is a parabola.

Philosopher. So if the ceiling were a large piece of film, a parabola can be regarded as a grossly-distorted picture of a circle?

Teacher. Exactly. And by turning the camera even further to get a hyperbola, the information has been so tremendously distorted that one has to think a bit to realize that it's just a different view of a circle.

Philosopher. So undistorted pictures of the circle are actually quite rare—you get them only when the film happens to be perpendicular to a circular cone's axis of symmetry.

Student. If being perpendicular is so good, why not just curl the film somehow, so the light rays always hit it at right angles?

Teacher. A camera with bent film?!

Student. Why not?

Philosopher. Maybe you have an idea here! I once visited a cinema where the room was circular, and there were projectors aimed at successive sections of the room. It gave a 360° view of everything. It was pretty impressive.

Teacher. You're saying the walls played the role of film bent into a cylinder? Interesting. But the light rays from the projector couldn't have hit every point of the screen orthogonally.

Philosopher. Come to think of it, we'd have perpendicularity on at most one circle of the cylindrical wall, wouldn't we?

Teacher. Yes. You know, in less than a minute I find myself changing from disbelief to now wondering about even *more* bending! A cylinder curves in only one direction, so to speak, and we need a room curving uniformly in *all* directions.

> The Euclidean plane often keeps us from seeing larger truths about conics.

Student. I once visited a planetarium. The whole place was the top half of a sphere.

Philosopher. Aha! Maybe we could let the screen—the film—be spherical. If the vertex of the circular cone—our light source—is at the center of the sphere, then no matter how we angle that source of light, we always record without distortion. If the cone is circular, we get two opposite circles. If the cone is elliptical, defined, say, by

$$z^2 = \frac{x^2}{2^2} + \frac{y^2}{1^2},$$

then we get two elongated curves on opposite sides of the sphere.

Teacher. I wonder if there exists a truly *ideal* screen... If there is such a screen, it should have no distortion, and record exactly one point for each double ray through the vertex. The plane not only creates distortion, it also fails to record light rays parallel to it. The sphere has no distortion, but as far as intersections are concerned, it's almost too much of a good thing—it records *two* points for every double ray!

Philosopher. Yes, I agree.

Teacher. However, I have an idea. Topologists have long used the trick of "identifying points"—for example, one can identify points to represent topological two-dimensional spaces such as a cylinder, Möbius strip, torus, sphere, the projective plane or a Klein bottle using a flat piece of paper, as in Figure 1.7.

I'd like to make a proposal:

Let our viewing screen be a sphere with opposite points identified.

With this definition of viewing screen, there's no distortion and every double ray intersects the screen in exactly one point.

Philosopher. This sounds like a promising idea.

Teacher. In a sphere with opposite points identified, any double cone intersects the sphere in two opposite curves, each pair of diametrically opposite (that is,"antipodal") points being identified.

Student. But we don't go around carrying globes with curves drawn on them! We use ordinary paper, like anybody else.

Teacher. I have to agree—all this is three-dimensional, and we need to describe exactly how to draw sphere-plus-curve on a piece of paper. We need an "official agreement" of some sort. Let me propose this:

- Use a unit sphere centered at $(0, 0, 0)$ in (x, y, z)-space;
- Place the cone's vertex at $(0, 0, 0)$;
- Vertically project the top half ($z \geq 0$) of the sphere to the (x, y)-plane. Of course, this projects any curve in the top hemisphere to the (x, y)-plane, as well.

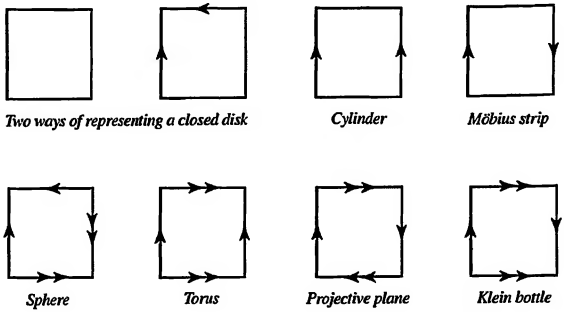


FIGURE 1.7. Various ways of identifying the sides of a square create an assortment of familiar topological two-dimensional spaces. Two sides are *identified* if they're marked with arrows of the same kind (both single-headed or both double-headed)—that is, we consider that the sides are glued together so that the arrow directions agree. Representing such two-dimensional topological spaces using identifications is useful because physically making such surfaces can be challenging or impossible in ordinary 3-D space. For example, at the easy end, to make a Möbius strip one need only "stretch the square," replacing it by a rectangle long enough so the paper won't tear when twisted. But neither the projective plane nor the Klein bottle can be realized in 3-D without introducing self-intersections. Notice that those surfaces in the top row all have some unmarked sides; after the gluing process, these unmarked sides remain, forming a boundary. On the other hand, all four sides of each figure in the bottom row have arrows, so each of these figures is without boundary. The Möbius strip, projective plane and Klein bottle are all nonorientable. For topological two-manifolds, this turns out to be equivalent to "one-sidedness," familiar in the Möbius strip. Each figure in the bottom row has a genus: from left to right it is 0, 1, 1, 2.

The image of the hemisphere in the (x, y)-plane is the "universe" in which we represent all conics, and any conic in that universe is the vertical projection of a curve on the sphere.

Philosopher. Your projection includes the hemisphere's equator, and its opposite points are identified. Does that mean the "universe" is actually a closed disk with opposite boundary points identified?

Teacher. Precisely! A closed disk with opposite boundary points identified is topologically identical to a sphere in which each pair of diametrically-opposite points gets identified. As yet another way of looking at it, one could add a point at infinity to the "end" of each 1-subspace of the (x, y)- plane and set up a topological equivalence between this extended plane and our disk.

Philosopher. Well, I'm looking at that assortment of two-dimensional topological spaces in Figure 1.7, and it appears that what you're describing is topologically the same as the pro-

jective plane. Although I know Figure 1.7 is a popular and traditional way of representing these spaces using identification, there's no reason why we couldn't use disks instead. The sides of the square could correspond, say, to the quarter-circles between the four compass points N, W, S, E.

Teacher. I agree.

> Artists struggled for centuries with the problem of perspective. Leonardo Da Vinci (1452–1519) dreamed of a mathematical theory for it, but that had to wait for Desargues' classical 1639 work on perspective (Gerard Desargues, 1591–1661). This highly original work was all but neglected, however; analytic geometry was the rage and had people's attention. On top of that, Desargues wrote in an impossibly arcane, eccentric style, filled with his own quirky nomenclature. All original copies were eventually lost, and it took another two centuries before the geometer Michel Chasles (1793–1880) chanced upon the work—a single copy made by one of Desargues' students.

Student. But I don't think we should totally ignore ordinary conics. Isn't there a relation between ordinary conics in $\mathbb{R}^2$ and the ones we're considering in the disk?

Teacher. Yes, and this picture illustrates the relationship:

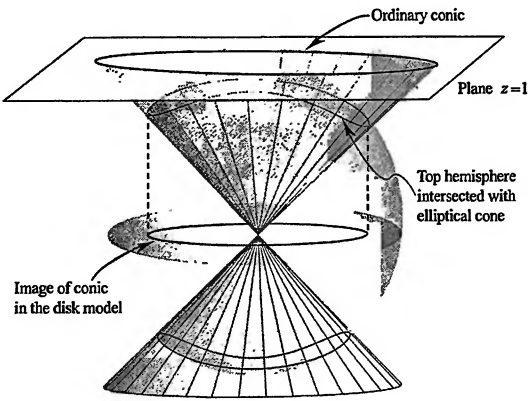


FIGURE 1.8. In the sphere model of the projective plane, opposite points of the unit sphere are identified. The ordinary conic is the central projection of [sphere] $\cap$ [elliptical cone] into the plane $z = 1$. The disk model conic is the orthogonal projection of [upper hemisphere] $\cap$ [elliptical cone] into the plane $z = 0$. More generally, one gets translated ellipses and hyperbolas by rotating the cone (and nothing else) from the vertical.

The illustration shows two plane projections. First, the vertical or *orthogonal* projection: the image of [upper hemisphere] ∩ [cone] lies in the disk bounded by the equator, and is a conic in the disk model. The other is a central projection: a light source at (0, 0, 0) casts a shadow of [sphere] ∩ [cone] on the plane $z = 1$. That shadow is the ordinary conic in which the cone cuts the viewing plane $z = 1$. To distinguish it from the projective plane, the ordinary viewing plane is often called the *affine plane*, and any curve in it, an *affine curve*.

Student. So the traditional universe is a plane, and the new one, a disk?

Teacher. Yes.

Student. It seems that the inverse of the central projection followed by the orthogonal one maps the entire plane into a disk!

Teacher. Actually, it maps the plane into the disk's *interior*. The disk's boundary points are extra, and can be thought of as coming from points at infinity added to the ordinary plane.

Student. It would probably be pretty complicated, but is there some function specifying exactly where a point on the plane ends up in the disk?

Teacher. There is, and it's actually quite simple. Let $(x, y, 1)$ be a typical point in the viewing plane $z = 1$. To get its central image on the top half of the unit sphere, just divide through by that point's distance from (0, 0, 0). That image is therefore

$$\left(\frac{x}{\sqrt{x^2+y^2+1}}, \frac{y}{\sqrt{x^2+y^2+1}}, \frac{z}{\sqrt{x^2+y^2+1}}\right).$$

The vertical projection simply drops the third component! So if you want to think of the whole plane as shrinking into the unit disk's interior, the function would be

$$(x, y) \to \left(\frac{x}{\sqrt{x^2+y^2+1}}, \frac{y}{\sqrt{x^2+y^2+1}}\right). \tag{1.1}$$

Notice that the image of (x, y) is a distance

$$\frac{\sqrt{x^2+y^2}}{\sqrt{x^2+y^2+1}}$$

from the origin. This distance is less than 1, and approaches 1 as (x, y) gets arbitrarily far away from (0, 0).

Philosopher. Interestingly, some 19th century artists secretly used a pinhole box or camera obscura to project an image of a scene on a drawing surface, which they'd then trace out. That surface is essentially the plane $z = 1$ in Figure 1.8. Figure 1.9 [p. 13] shows the setup more explicitly.

Philosopher. I've been thinking about Figure 1.8, and I find myself struggling to see some essential "oneness" to all the viewing screens. There are actually four so far—the tangent plane, the equatorial disk, the upper hemisphere, the entire sphere—and we probably could come up with others, too. I guess I'm looking for some one idea that we can point to and say, "Look—all those others are just special cases..."

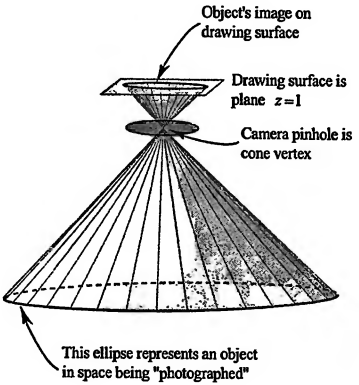

Object's image on drawing surface
Drawing surface is plane $z=1$
Camera pinhole is cone vertex
This ellipse represents an object in space being "photographed"

FIGURE 1.9. The object being sketched or photographed emits or reflects light rays. Those rays that happen to pass through the pinhole form a cone—the cone appearing here, with vertex at the pinhole. The cone's (light) rays continue on past the pinhole until they hit the drawing surface.

Teacher. Perhaps I can put some perspective on things. It seems to me that the closest thing to a unifying idea would be a double light ray—*a whole line through the vertex of the cone*. I say that because all we've done is just intersect that line with various objects like a plane or sphere. If an object intersects all such lines in a reasonable way, we like it and say, "Hey, here's a viewing screen!"

Student. What do you mean, "reasonable"?!

Teacher. Basically, if the object always intersects lines in *points*, we like that, and it seems reasonable to us, mainly because we human beings have developed a culture—an ingrained habit—of thinking of points as little dots or flyspecks. They're easy to visualize, and the idea goes back thousands of years. Yet Euclid was careful, and treated a "point" as undefined. He never said, for example, that a point looks like a dot. In fact, diagrams are just aids, concessions to our human need to have something visual to help us along in carrying out proofs, using logic together with a set of axioms. A computer or robot might automatically churn out little proofs of things in Euclidean geometry, and *it* has no visual conception of anything!

Philosopher. I once had a far-out teacher who said that you could do Euclidean geometry thinking of the points as *vertical lines* through a horizontal plane! A line would then look like a vertical plane, a circle would look like a vertical cylinder, and so on. He said that you could visualize everything that way, and it wouldn't affect any of the arguments in Euclidean geometry since, as you said, "point" is undefined. How we conceptualize "point" to make us functional mathematicians, is our business. Who am I to say how a person, blind from birth, might conceptualize "point"?

Teacher. Everything you say sounds right. In place of your teacher's vertical lines, you could use, say, two dots: just draw a parallel plane one inch above your bottom plane, and let everything have a double. You could call two flyspecks, one right above the other, a "point." There's nothing sacred about always having to think of a point as a single dot.

So, I'm thinking this: forget about looking at points as dots. Let them be lines through a common vertex. After all, it's those lines that actually generate most of the pictures we have been looking at, so let those lines be the ultimate "points"!

Student. This sounds crazy! You're seriously calling a line, a *point*?

Teacher. Yes.

Student. But why are we doing all this? I thought we already agreed that the sphere with opposite points identified was an "ultimate viewing screen"!

Teacher. Well, the disk, hemisphere and sphere models are all "ultimate," in that they're all essentially the same. But for us human beings, each has a slightly different advantage: the disk model is flat, and we clearly see points at infinity—the disk model is good for drawing pictures. The hemisphere removes the distortion in the disk model, but we lose the easy-to-draw flatness. The sphere doesn't ever visually break up loops, and is very symmetric, so that's nice, too.

Student. But, again, why are we calling a line, a *point*?!

Teacher. Because doing that brings all three of our "ultimate" screens under one umbrella. That is, in each of the disk and hemisphere models, there's a one-to-one correspondence between its points and lines through the vertex. In the sphere model, there's a one-to-one correspondence between these lines and diametrically opposite—"antipodal"— point-pairs. So all I'm saying is: instead of always dealing with various flavors of *representatives*, deal with the *lines themselves*.

Student. I think I see what you're saying.

Teacher. In fact, I'd like to go a step further, and propose that we take the origin of 3-space to be that common vertex.

Definition 1.1. Real projective n-space, denoted $\mathbb{P}^n(\mathbb{R})$, consists of the 1-spaces of $\mathbb{R}^{n+1}$

Philosopher. Any particular reason why?

Teacher. Because then we've linked up this whole business with vector spaces—the 1-subspaces of a vector space are precisely the lines through the origin. With this link, you suddenly inherit many useful concepts and facts from a well-established and developed field, linear algebra.

Philosopher. Ingenious! I also sense a tremendous potential here. Maybe I'm stretching things a bit, but it seems reminiscent of Descartes' linking up Euclidean geometry and algebra. That simple act harnessed the power of algebra to help prove, suggest and conceptualize things in geometry. So here, as I see it, you're linking up a more modern kind of algebra—linear algebra—with another part of geometry, where points at infinity are important.

Teacher. I see your analogy.

Here are two alternative ways of looking at the points of $\mathbb{P}^n(\mathbb{R})$:

- They are the points of an n-sphere in $\mathbb{R}^{n+1}$ in which diametrically-opposite points are identified.
- They are the points of a closed n-ball in $\mathbb{R}^n$ whose diametrically-opposite boundary points are identified.

The first of these is the **sphere model** of $\mathbb{P}^n(\mathbb{R})$.

The second is the **ball model** of $\mathbb{P}^n(\mathbb{R})$, also called the **disk model** when $n = 2$.

Student. So you're thinking of points as 1-subspaces of $\mathbb{R}^3$, and you say facts about vector spaces can help us. Well, I definitely need some help! Can you give me a specific example of what you mean?

Teacher. Yes. We can make literal translations from vector space facts to projective geometry facts.

Definition 1.2. Let $\mathbb{R}^{k+1}$ be a $(k+1)$-dimensional subspace of $\mathbb{R}^{n+1}$. The set of all 1-spaces of $\mathbb{R}^{k+1}$ defines a **projective subspace of** $\mathbb{P}^n(\mathbb{R})$, which we denote $\mathbb{P}^k(\mathbb{R})$. A subspace $\mathbb{R}^n$ of $\mathbb{R}^{n+1}$ thus defines a projective subspace $\mathbb{P}^{n-1}(\mathbb{R})$ of $\mathbb{P}^n(\mathbb{R})$. We may select one particular such subspace and call it the **(projective) hyperplane at infinity of** $\mathbb{P}^n(\mathbb{R})$. The points of this $\mathbb{P}^{n-1}(\mathbb{R})$ are then **the points at infinity of** $\mathbb{P}^n(\mathbb{R})$. The remaining "finite" points of $\mathbb{P}^n(\mathbb{R})$ are those 1-subspaces of $\mathbb{R}^{n+1}$ that are not in $\mathbb{R}^n$; they form **affine n-space.** It can be shown that a parallel copy of $\mathbb{R}^n$ intersects each 1-subspace of $\mathbb{R}^{n+1} \setminus \mathbb{R}^n$ in a single point; this is called a **vector space model of affine n-space.** $\mathbb{P}^2(\mathbb{R})$'s hyperplane at infinity is called **the line at infinity.**

Here are two alternative ways of looking at the hyperplane at infinity and its points:

- In the **sphere model** of $\mathbb{P}^n(\mathbb{R})$, any hyperplane $\mathbb{R}^n$ of $\mathbb{R}^{n+1}$ through the sphere's center intersects the sphere in a "great sub-sphere," and we may take any particular such subspace to be $\mathbb{P}^n(\mathbb{R})$'s hyperplane at infinity, its diametrically-opposite points being identified. When $n = 2$, this means that in an ordinary sphere, any of its great circles, with its diametrically-opposite points identified, may be taken as the line at infinity.
- In the **ball model** of $\mathbb{P}^n(\mathbb{R})$, the boundary of the ball is an $(n-1)$-sphere in $\mathbb{R}^n$, and this sphere, with its opposite points identified, may be taken as the hyperplane at infinity. The open ball is then the **ball model of affine n-space.** When $n = 2$, the ball is a disk, its boundary, a circle; then the hyperplane at infinity is this circle with its diametrically-opposite points identified; the open disk is the **disk model of affine 2-space.**

In the Euclidean plane, one can study properties of figures invariant under Euclidean transformations. Analogously, in $\mathbb{P}^n(\mathbb{R})$, one can study properties of figures invariant under "projective transformations."

Definition 1.3. A nonsingular linear transformation on $\mathbb{R}^{n+1}$ maps 1-subspaces to 1-subspaces, and therefore maps points to points in $\mathbb{P}^n(\mathbb{R})$. These maps on $\mathbb{P}^n(\mathbb{R})$ are called **projective transformations of** $\mathbb{P}^n(\mathbb{R})$.

Definition 1.4. In $\mathbb{R}^n$, a nonsingular linear transformation followed by a translation is called an **affine transformation** of $\mathbb{R}^n$. Two subsets connected by an affine transformation are said to be **affinely equivalent.**

Intuitively, affine transformations are those projective transformations sending the hyperplane at infinity to itself.

Paradigm example: Let the 1-spaces of the (x, y)-plane in (x, y, z)-space $\mathbb{R}^3$ define the points of $\mathbb{P}^2(\mathbb{R})$'s line at infinity. The plane $z = 1$ then intersects any other 1-space in a point, and thus may be taken as a vector space model of an affine plane. Its affine transformations are induced by linear transformations of $\mathbb{R}^3$ having the form

$$(x, y, z) \to (x, y, z)\begin{pmatrix} a & b & 0 \\ c & d & 0 \\ e & f & 1 \end{pmatrix},$$

with $ad - bc \neq 0$. These send the (x, y)-plane to itself, and when multiplied out, we see that in the plane $z = 1$, they define a nonsingular linear transformation followed by a translation.

Some Examples

Student. Suppose you perform a general 3-D rotation of the cone about its vertex, which we've taken to be the origin. The cone always intersects the sphere, so you get a corresponding rotating curve on the sphere. If you orthogonally project from the upper hemisphere into the disk, you get a changing curve in the disk model ...

Teacher. That's right.

Definition 1.5. The union of any collection of 1-spaces of $\mathbb{R}^n$ is called **homogeneous.** Let S be any set in $\mathbb{R}^n$. The union of all the 1-subspaces of $\mathbb{R}^n$ through points of S is called the **homogenization** of S, and is denoted $H(S)$. The operation $S \to H(S)$ is called **homogenizing** S. We call the topological closure in $\mathbb{R}^n$ of $H(S)$ the **closed homogenization** of S, and denote it $\overline{H}(S)$.

We always include the word "closed" or the overbar if we mean closed homogenization.

Both $H(S)$ and $\overline{H}(S)$ are homogeneous. Visually, the homogenization or closed homogenization of an object is a generalized "cone with vertex at the origin." Figure 1.8 on p. 11 shows the cone in $\mathbb{R}^3$ defined by an ellipse. ***Closed*** homogenization is useful because if the plane curve in either figure is defined by a polynomial, then the closure of the cone containing it is also defined by a polynomial—an important analytic tool. (The locus, or "zero set," of any polynomial is closed.) In Figure 1.8, ordinary homogenization always omits any line contained in the horizontal plane through the vertex.

Student. I sure could use some examples so I can see things better!

Teacher. Well, let's make a movie! I suggest we start with an ellipse drawn on the ordinary viewing screen $z = 1$. Its homogenization is an elliptical cone in $\mathbb{R}^3$. We can then rotate this cone about the origin and vertically project [upper hemisphere] $\cap$ [cone] to make a movie in the unit disk.

There are a couple of tricks to simplify this. First, ink the curve [sphere] $\cap$ [cone] on a white sphere, throw away the cone, and just rotate the sphere. Second, the vertical projection of the upper hemisphere into the (x, y)-plane is very nearly the same as *assuming the inked sphere is viewed from a relatively great distance*. An example would be looking at craters on the moon. Its distance from us is relatively great in the sense that it's many moon-diameters away—but of course not so many that the moon looks like a point!

Student. Since a hemisphere is all you see at once anyway, doesn't that mean we just rotate the sphere and view the spectacle from way off?

Definition 1.6. The natural topology on $\mathbb{P}^n(\mathbb{R})$ has as its **open sets** the (ordinary) homogenizations of the open sets of $\mathbb{R}^{n+1}$

Teacher. Yes, that's exactly the idea. Visually, you can't tell the difference between that and the mathematically true orthogonal projection.

Student. That seems simple enough!

Teacher. Let's begin making our movie. To make it a bit more informative, let's also include the four lines tangent to the ellipse that define the fundamental rectangle. Figure 1.10 shows a picture of all this in the plane $z = 1$.

The ellipse defines an elliptical cone through $(0, 0, 0)$, and its intersection with the sphere is a "spherical curve"—the curve bending somewhat, due to the sphere's curvature. Each of the four lines defines, *via* closed homogenization, a *plane* through $(0, 0, 0)$. Each of these four planes intersects the sphere in a great circle. Figure 1.11 shows what all this looks like from a distance.

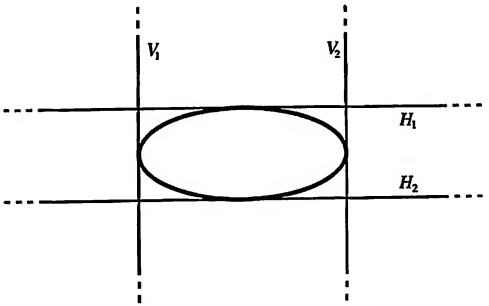


FIGURE 1.10. This is the conic in the plane $z = 1$ of Figure 1.8 on p. 11, together with its four principal tangent lines.

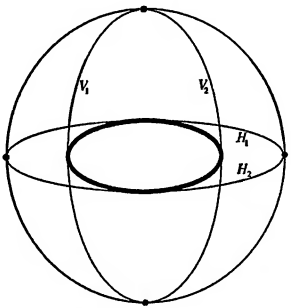


FIGURE 1.11. The ellipse is the vertical projection of the [upper hemisphere] $\cap$ [cone] of Figure 1.8. Also shown are the four tangent lines sketched in Figure 1.10. Here, they appear as curves.

The orthogonal projection of any circle (great or not) is an ellipse. Therefore, because we see the projection of only a half-sphere, H_1, H_2, V_1, V_2 are all half-ellipses. But in Figure 1.11, H_1 and H_2 are symmetric with respect to the disk center, so $H_1 \cup H_2$ makes up a full ellipse tangent to the disk's boundary at diametrically-opposite points. Similarly for $V_1 \cup V_2$.

We create our "movie" by using a number of not-too-large increments.

> When viewed from a relatively great distance, sphere rotations nicely approximate projective transformations of a disk model of $\mathbb{P}^2(\mathbb{R})$.

Figure 1.12 shows the inked sphere in Figure 1.11 after an approximately $\frac{\pi}{8}$ turn, and here we already see some nonuniform distortion. The left side of the ellipse is slightly more blunted, for example.

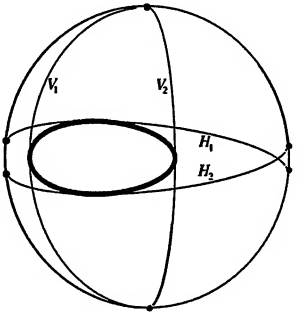


FIGURE 1.12. Rotating the sphere with the ellipse drawn on it makes the ellipse look slightly asymmetrical. Also sketched here are the ellipse's four principal tangent lines. The closed homogenization $\overline{H}(L)$ of any line L in the affine viewing plane $z = 1$ is just a 2-subspace of $\mathbb{R}^3$. This 2-subspace intersects the sphere in a circle, and we see half of the circle. In this disk model, that semicircle looks like a half-ellipse.

Let's rotate the sphere so that the leftmost part of the loop is tangent to the great-circle boundary. This is shown in Figure 1.13.

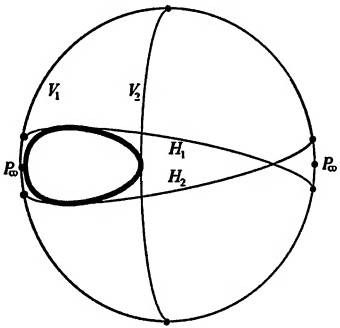


FIGURE 1.13. The ellipse in Figure 1.12 has been additionally rotated so it appears tangent to the disk's boundary. This corresponds to a parabola in the ordinary plane. This picture therefore shows a disk view of a parabola, and its loop nature is easy to see. The isolated point P_∞ on the right is the same point as the one opposite it on the left.

The leftmost point of the circle is now exactly on the left "horizon," and simultaneously, a point opposite it appears on the right horizon. In (x, y, z)-space, the line containing these two points is a double light ray emanating from the origin. It lies, of course, in the (x, y)-plane and is parallel to the ordinary viewing plane $z = 1$. It never touches this plane, but intersects it in an added point at infinity, which is what we see at the boundary in our disk picture. Of course, we can look again at the light source—a double cone—and realize that it, too, has rotated. This rotated cone intersects the plane $z = 1$ in a parabola opening to the left.

> The earth's moon has a slight bulge, and the extra gravitational pull is why it always faces us. But not precisely: it slowly oscillates some 9° to our left and right. In this 18° nutation, our view of craters experiences a continual close-to-the-identity projective transformation.

Here's a picture of our inked-on ellipse after the sphere has been given an additional rotation:

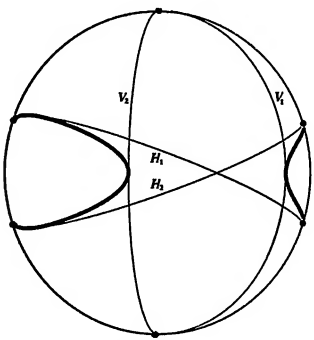


FIGURE 1.14. Some points of the ellipse have rotated out of view, but on the right, corresponding points of the antipodal ellipse are coming into view. What we see as separate branches actually form one connected loop, because antipodal points are identified or "glued together."

At this stage, we now see substantial parts of each of the two opposite ellipses. Agreeing that antipodal points of the sphere are topologically identified, we're not actually looking at parts of different ellipses, but at different parts of the same ellipse. The rotated double cone now intersects the plane $z = 1$ in a hyperbola, and the left and right parts of the ellipse correspond to the left and right branches of the hyperbola.

Philosopher. This is great, because it answers a problem about unification that I hadn't even mentioned: "If conics are so unified, then how come an ellipse has just one connected

piece, while a hyperbola has two?" This answers that question by exposing a substantial drawback of a flat screen: it can miss recording those points "at infinity." These missing points create breaks in what ought to be a connected loop. The sphere or disk model catches them all, and in addition keeps everything bounded so we can see what's going on at infinity. That makes it easy to traverse the loop all the way around in all cases. The plane model's lack of faithfulness gives the wrong impression. It is remarkable how simply spinning a sphere with two opposite ellipses printed on it can so easily help unify ellipses and hyperbolas!

Teacher. I agree.

> Using a plane as a viewing screen can create breaks in topological loops, artificially separating them into two parts, which we call hyperbola branches.

Here's what we see after a full $\frac{\pi}{2}$ rotation:

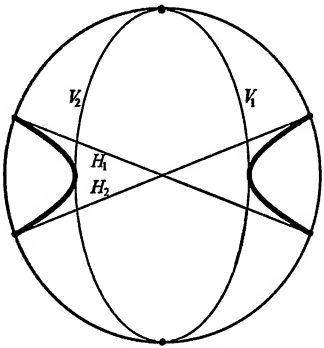


FIGURE 1.15. The sphere rotated $\frac{\pi}{2}$. The original front-and-center mug shot of an ellipse has by now morphed all the way into a front-and-center shot of a hyperbola. Importantly, it also shows what the asymptotes to this hyperbola really are: just the horizontal tangents to an ellipse viewed from a different perspective.

Interactive Applets

For an interactive arbitrarily rotatable sphere generalizing Figures 1.11–1.15, see Chapter Applets 1 and 2 on the CD-ROM.

The hyperbola branches appear symmetrically. This picture answers one of the big questions raised earlier. If we go backwards from Figure 1.15 to Figure 1.11, we see ex-

actly what becomes of hyperbola asymptotes—*they morph into principal tangent lines to the ellipse*! The flat plane ends up keeping the word "asymptote" disconnected from what it really is: a tangent line whose point of tangency happens to be at infinity. Since the choice of line at infinity is just a matter of choosing a viewing perspective, we see that whether we say "asymptote" or "tangent line" is only a matter of our viewing position.

The term *asymptote* comes from the Greek "a" (*not*), "sym" (*together*), and "ptote" (*to fall*). Roughly, an asymptote "doesn't fall together"—it never quite touches the curve it approaches.

One can concretely see pictures corresponding to further rotations, by turning this book upside down, interchanging V-labels and reading from Figure 1.14 down to Figure 1.11. Beyond that, the sequence simply repeats itself.

Jean-Victor Poncelet
1788–1867

It was Poncelet who gave projective geometry its standing as an important geometry in its own right. As a young soldier in his mid-twenties, he was captured in Napoleon's disastrous Russian campaign and as a prisoner of war, forced to march some five months to solitary confinement in Saratoff on the Volga River. In his cell, with no books, he mused endlessly about his beloved geometry, trying to remove exceptions, and slowly bringing into focus fundamental ideas of projective geometry. Upon returning to France after the war, he wrote up his conclusions in the highly original (and controversial) *Traité des propriétés projectives des figures*—a work that attracted many young students to him. Thus began the "great period" of projective geometry.

Chapter Summary

Viewing conics exclusively in the real plane hides much of their underlying unity. An important first step is to add a point at infinity to each line, with all mutually parallel lines sharing the same point. The set of all these added points forms a new line L_∞, the "line at infinity," and the original real plane together with L_∞ forms the real projective plane $\mathbb{P}^2(\mathbb{R})$. In it, any two distinct lines intersect in a point. When we look at conics in this extended "universe," we find that every ellipse, parabola or hyperbola is a topological loop. In the case of a hyperbola, for example, its familiar branches are actually two halves of such a loop. Figures 1.11–1.15 (and even more informatively, Chapter Applets 1 and 2 on the CD-ROM) summarize the main point of this chapter: an ellipse simply changes into a parabola or hyperbola as one's viewing perspective changes. We find that certain parallel tangent-line pairs naturally morph into the asymptotes of a hyperbola.

Exercises

1. (a) Hold a penny directly under a point light-source illuminating a table top, thereby creating an elliptical shadow. Since its shape varies as you turn the penny this way or that, some people think this illustrates the plane-slicing-cone idea. But you can never get hyperbolas this way. *Why not*?

   (b) Replace the light bulb by the sun on a cloudless day, allowing us to assume the incoming light rays are parallel. Does this change any aspect of (a)?

   (c) In (a), can you get hyperbolas by moving the penny so it isn't directly under the light?

   (d) In practice, you need to hold the penny fairly close to the table top. Why? Is the same true when the sun is used? Would it make any difference if you were on the moon, where there's no atmosphere?

2. (a) It's a common misconception that if you take a circular cone and slice it with various planes not going through the vertex, then you can get every shape of nondegenerate conic. You *do* get every possible ellipse this way, but not every hyperbola. What, exactly, is the problem? What specific hyperbolas are excluded?

   (b) What cone could you use to get, say, the hyperbola

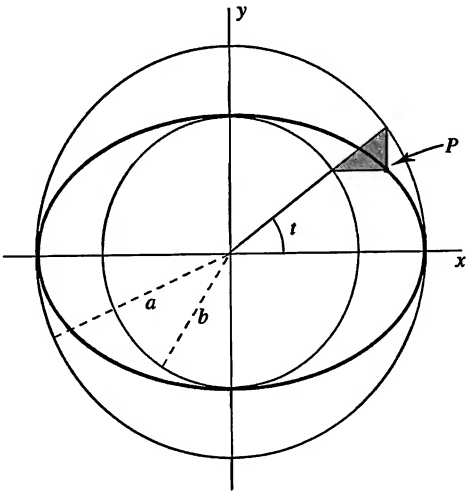
$$\frac{x^2}{100^2} - \frac{y^2}{1^2} = 1?$$

3. It is believed that looking at the shadows of circles inspired the ancients' first contact with the ellipse, probably by Menaechmus (ca. 350 BCE). This represents one explicit, geometric approach among today's wide panoply of approaches ranging from geometric to algebraic to analytic. At the analytic end are parametric equations such as $x = a\cos t$, $y = b\sin t$. Here, too, one finds a common misconception: some people assume that t is the angle from the origin to the point $(a\cos t, b\sin t)$. A quick sketch shows this is true only when $a = b$—that is, when the ellipse is a circle. What, then, *does* the angle t represent? Let's find out.

(a) Using $\cos^2\theta + \sin^2\theta = 1$, prove that $x = a\cos t$, $y = b\sin t$ parametrizes

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1.$$

(b) Draw two concentric circles of radii a and b, where $a > b$:



Note that for any t, the shaded right triangle has hypotenuse $a - b$. Show that the coordinates of P are $(a\cos t, b\sin t)$. Thus the heavily-drawn curve is in fact the ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1,$$

and we can see what t's actual role is.

(c) The picture shows that in the first quadrant, the angle from the origin to P lags behind t. The lag is the greatest when $t = \frac{\pi}{4}$, and is 0 when t is 0 or $\frac{\pi}{2}$. In the second quadrant, the "lag" is negative. Sketch the signed lag as a function of t, $0 \le t \le 2\pi$.

(d) In the diagram, P is the shaded triangle's vertex lying *below* its hypotenuse. Philosopher could well wonder: Why *below*? Why not *above*? As so often happens, his symmetry-seeking radar is on to something. What ellipse, and what parametrization, does the "above" approach lead to?

4. In the disk model of $\mathbb{P}^2(\mathbb{R})$, the metric becomes more "concentrated" as one moves toward the disk's boundary. A **metric** $d(P, Q)$ is a nonnegative function d on point-pairs (P, Q) of a space, with d satisfying: $d(P, Q) = 0$ if and only if $P = Q$;

$d(P, Q) = d(Q, P)$; and $d(P, R) \leq d(P, Q) + d(Q, R)$. Now try to define "metric concentration," and using your definition, plot a graph of metric concentration versus r, for $0 \leq r < 1$. [Hint: The map in (1.1) on p. 12 has an inverse which sends each point of the open disk to a point of the ordinary plane. Use this to define $d(P, Q)$ for points P, Q in the open disk.]

5. Show that the Klein bottle in Figure 1.7 (p. 10) is not orientable by appropriately pushing around a small oriented test loop on it. You should be able to get the loop to come back to its starting position with its arrow reversed.

6. Let T be any affine transformation of the affine plane $z = 1$ in (x, y, z)-space. Prove that in that plane, if lines L_1 and L_2 are parallel, then so are $T(L_1)$ and $T(L_2)$.

2

Life at Infinity

in which the trio travels to infinity and looks around

The Line Euclid Missed

Philosopher. The more I think about the mapping

$$(x, y) \to \left(\frac{x}{\sqrt{x^2+y^2+1}}, \frac{y}{\sqrt{x^2+y^2+1}} \right)$$

in (1.1) [p. 12], the more remarkable it seems. It symmetrically shrinks the entire plane down to an open disk. That's powerful, because in one stroke, it converts something sounding far-fetched (adding points at infinity to a Euclidean plane), to something entirely routine (adding boundary points to an open disk). It compresses the entire Euclidean plane so much that one can just sit and contemplate the whole thing—and then add in the boundary of this universe to boot. That's awesome.

Teacher. It sure is.

Philosopher. This map (1.1) seems to shrink things ever more strongly the further we get from the origin.

Teacher. That's right. Uniform shrinking wouldn't work at all—for example, an intensely contracting *linear* map would do nothing of this sort. The metric in the disk gets more and more concentrated toward the boundary.

Student. What is it right *on* the boundary?

Teacher. Actually, we've transported the metric on the Euclidean plane only into the *open* disk. We then elected to add boundary points to the disk because they seem to correspond naturally to points infinitely far away. But *the metric itself doesn't extend to the boundary.* You'd have to say the distance from one boundary point to any other point of the closed disk is "infinity," which is not a real number. A metric function assumes only real, and therefore finite, values.

Philosopher. I sense that because the metric gets more concentrated toward the edges of the disk, then in the disk, the images of two parallel lines in the plane actually approach each other the further out you go...

Teacher. Yes. In fact, we can use the mapping (1.1) to find precisely what these images look like. Since the shrinking is symmetric about the origin, we can get the idea by looking at the image of a horizontal line, say the one consisting of all points (t, a), with fixed $a \neq 0$. Under the map in (1.1), (t, a) goes to

$$\left(\frac{t}{\sqrt{t^2+a^2+1}}, \frac{a}{\sqrt{t^2+a^2+1}}\right).$$

Setting the first and second coordinates equal respectively to x and y and then squaring, gives

$$x^2 = \frac{t^2}{t^2+a^2+1},$$
$$y^2 = \frac{a^2}{t^2+a^2+1}.$$

We can eliminate the t's by choosing an appropriate linear combination of x^2 and y^2:

$$1 \cdot x^2 + \left(\frac{a^2+1}{a^2}\right) \cdot y^2 = \frac{t^2+a^2+1}{t^2+a^2+1} = 1.$$

Student. But

$$x^2 + \frac{a^2+1}{a^2} y^2 = 1$$

is just an ellipse!

Teacher. Right—it has semi-major axis 1 and semi-minor axis

$$\frac{|a|}{\sqrt{a^2+1}} < 1.$$

The actual image is the top or bottom half of this ellipse. And when $a = 0$, the image is just the horizontal diameter of the disk. As $|a|$ increases in size, the vertical dimension of the ellipse increases, tending toward, but never quite reaching, the bounding circle of the disk. The ellipse touches opposite boundary points of the disk.

Philosopher. We observed an instance of this in Figure 1.11, [p. 18], didn't we?

Teacher. Yes, and all this makes sense, geometrically. Start with any line L in the plane $z = 1$ in Figure 1.8 [p. 11]. The closed homogenization of L is a 2-subspace of $\mathbb{R}^3$, and this plane intersects the sphere in a great circle. The orthogonal projection of this circle is an ellipse or a line segment, in either case touching the disk in opposite boundary points. Figure 2.1 shows the images of a few lines chosen in Figure 1.8's plane $z = 1$, including four parallel ones.

Philosopher. Our idea of using the disk model of the projective plane as a screen seems to be paying some dividends! Figure 2.1 vividly shows lines intersecting at infinity. This answers a problem that has bothered me for years. I was always unhappy with the "exception" part in the famous Euclidean *dictum*, "Two distinct lines intersect in one point, *except* when the lines are parallel." My philosophy is that one should try to remove excep-

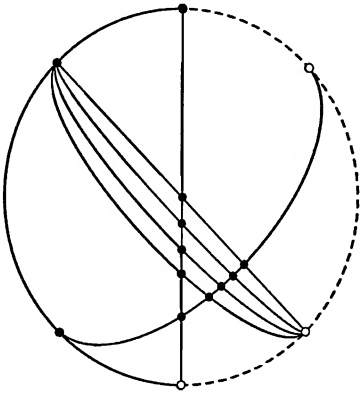


FIGURE 2.1. Seven lines in the projective plane. One of them is the bounding circle—the line at infinity. This picture illustrates that any two different lines intersect in exactly one point, even if one of those lines happens to be the line at infinity.

tions whenever possible. By adding points at infinity as in Figure 2.1, I see how and where parallel lines intersect in one point.

> The projective plane is highly symmetric: Any two distinct points determine a line, and any two distinct lines determine a point. This property can be included in a "self-dual" set of axioms for the projective plane: interchanging the words "point" and "line" in one axiom gives a dual axiom. Any theorem about points and lines, being proved *via* a string of references to these axioms, therefore has a dual, obtained by interchanging the two words. This "principle of duality" was used over and over again by Poncelet in his groundbreaking work *Traité des propriétés projectives des figures.*

Teacher. This isn't the only time that extending our mathematical world has removed exceptions! Before complex numbers were introduced, people could solve many polynomial equations, but there were also many others that seemed inherently "exceptional," such as $x^2 + 1 = 0$. Extending $\mathbb{R}$ to $\mathbb{C}$ removed those exceptions. It actually did more than that, because one could let even the polynomial *coefficients* be complex, and still be guaranteed solutions.

Philosopher. Do we get a similar bonus by adding the points at infinity?

Teacher. Yes, we do. All those extra points in $\mathbb{P}^2(\mathbb{R})$ themselves form a line—the line at infinity. Just as one gets solutions for an extended class of polynomials, one gets points of

intersection for the extended class of lines. Figure 2.1 suggests why this is so: not only do any two "finite" lines intersect in one point, but also the new line at infinity intersects any other line in one point.

Student. So you could say that this line at infinity is "the line Euclid missed"!

Teacher. A good description. For many centuries, nobody thought of compressing the whole plane down to an easy-to-visualize disk. But it greatly increases our ability to understand what is going on at infinity.

Any $\mathbb{R}^n$ can be compressed to an open n-ball using a formula like (1.1) on p. 12. Our recipe of adding boundary points and identifying opposite ones creates a $\mathbb{P}^{(n-1)}(\mathbb{R})$ at infinity, as in Definition 1.2 on p. 15. Example:

$$x \to \frac{x}{\sqrt{x^2+1}}$$

compresses $\mathbb{R}$ to an open interval. Adding the two endpoints and identifying them produces $\mathbb{P}^1(\mathbb{R})$—a projective line—which is topologically just a circle. Similarly,

$$(x,y,z) \to \frac{1}{\sqrt{x^2+y^2+z^2+1}} \cdot (x,y,z)$$

compresses $\mathbb{R}^3$ to an open ball. Adding boundary points and identifying pairs of opposites produces the ball model of $\mathbb{P}^3(\mathbb{R})$. The points at infinity of $\mathbb{P}^3(\mathbb{R})$ form $\mathbb{P}^2(\mathbb{R})$, just as the points at infinity of $\mathbb{P}^2(\mathbb{R})$ form $\mathbb{P}^1(\mathbb{R})$.

Is the Line at Infinity So Different?

Philosopher. The projective plane can be modeled by a sphere with opposite points identified. That's beautiful and symmetric. But in the *disk* model of the projective plane, the line at infinity is different from the other ones—for example, distance isn't defined on it. One is so nice and symmetric, while the other is so unsymmetric!

Teacher. The disk and the sphere with opposite points identified are both models of the real projective plane, but it happens that the disk model has some structure above and beyond what the projective plane requires. A projective plane needn't have a metric defined on it or on part of it, nor does there have to be a particular "line at infinity" chosen in a projective plane—these are extras that appear in the disk model. For example, the rotations depicted in Figures 1.11–1.15 [starting on p. 18] amount to choosing different perspectives. Instead of remaining stationary and observing a rotating, inked sphere, we could move around a fixed sphere. Of course the sphere's "horizon" is constantly changing as we move. The horizon, with its opposite points identified, serves as a "moving line at infinity," since the horizon is relative to a particular vantage point. In any such view, the non-horizon points of the sphere inherit a metric in the way we've just described.

Philosopher. So the distance between points on the sphere, as well as what we call the line at infinity, are *relative* to the viewing direction...

Teacher. Absolutely.

Student. Since the sphere with opposite points identified is so nice, why bother with a metric and a line at infinity? That seems a lot more complicated.

Teacher. True. But without it, all you can get are *qualitative* results. Choosing a specific screen allows us to write that screen's equation. For example, in Figure 1.8, the screen has equation $z = 1$, and that leads to the mapping (1.1) into the disk. This in turn leads to the equations used in plotting those curves shown in Figures 1.11–1.15, and without those equations, one might for example miss the bull-nose shape of the curve in Figure 1.13 [p. 19]. If I'm puzzled or curious about what's going on at some point at infinity, I can reposition myself to look straight down on that point at infinity, and work out new equations describing precisely what's happening there.

How to See Things at Infinity

Philosopher. That sounds pretty powerful. I'd like to see this at work in a specific example. Figure 2.1 shows some parallel lines meeting at infinity, and we found these converging curves to be half-ellipses in the disk. Now, *work your magic! Can you reposition me so I look directly down at the point at infinity where they all meet?* Can you get their equations?

Student. This is exciting.

Teacher. Yes, we can do that. Let's use a very simple set of lines, say the four horizontal lines L_n ($n = 1, 2, 3, 4$) in the (x, y)-plane having equations $y = n$. They look like this:

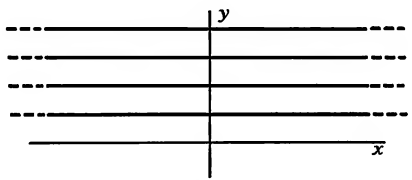


FIGURE 2.2. Travel right or left all the way to the point at infinity, and look around. What do these four horizontal lines look like there?

Let's consider this page to be the plane $z = 1$, tangent to the unit sphere. Visually, we're looking down the z-axis. We're now going to use precisely the spirit of Figure 1.1. Think of C as a *generalized cone*, or *homogeneous set* in $\mathbb{R}^3$—a set consisting of lines through $(0, 0, 0)$ in real (x, y, z)-space. The closed homogenization of any one of our four lines is a 2-space of this $\mathbb{R}^3$. You can see this in the next illustration, Figure 2.3.

Now when we used the double flashlight as in Figure 1.6 [p. 8] to get conics, the double cone moved, the ceiling being fixed. But the cone could equally well stay put and the screen could move. Figure 2.4 illustrates this: the cone is the same 2-subspace appearing in Figure 2.3, but now the screen, given by the equation $x = 1$, is vertical.

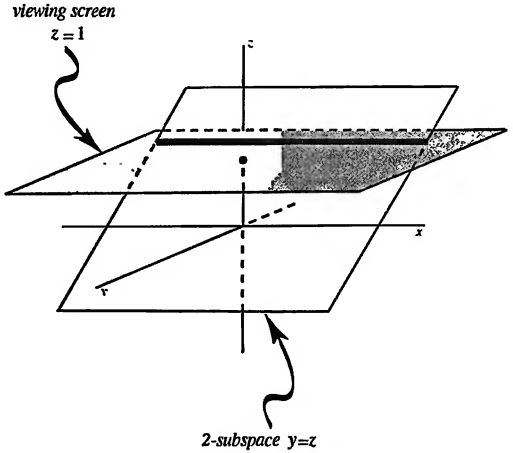


FIGURE 2.3. The horizontal, more heavily-shaded plane is the ordinary viewing plane—the plane of the paper Figure 2.2 is drawn on. The parent cone is in this case the lightly-shaded plane $y = z$, and is the closed homogenization of the line $y = 1$ in Figure 2.2.

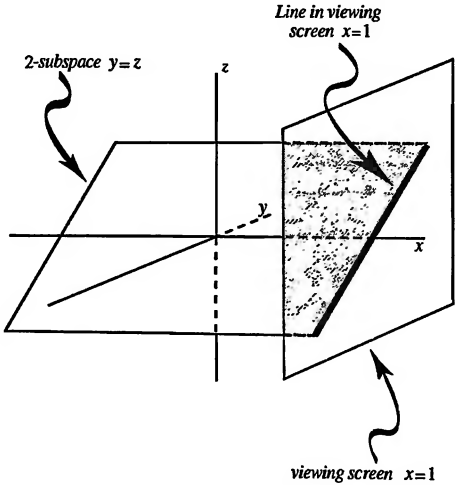


FIGURE 2.4. Here we see the parent cone $y = z$ in Figure 2.3 now intersected by a new viewing screen, $x = 1$. *This* screen allows us to view life at the "end" of Figure 2.2's x-axis. At that point at infinity, the line $y = 1$ in Figure 2.2 becomes the heavy line appearing here.

Of course changing the screen changes the picture it records, and in Figure 2.4 we see that the image in the plane $x = 1$ is now the heavily-drawn line there.

Here is an algebraic counterpart to the notion of *closed homogenization* introduced in Definition 1.5:

Definition 2.1. A polynomial is said to be **homogeneous of degree** m if the total degree of each term is m. If p is any polynomial of degree m in $\mathbb{R}[x_1, \ldots, x_{n-1}]$, then the **homogenization** of p, denoted $H(p)$, is obtained by multiplying each term of p by an appropriate power of a new variable (say, x_n) so that each term then has degree m. An **equation** $p = 0$ is **homogenized** by homogenizing p, and then the **equation is homogeneous.**

Example: If $p = y + xy - x^3 + x^2 + 1$, then $H(p) = yz^2 + xyz - x^3 + x^2z + z^3$

Fact: The locus in $\mathbb{R}^n$ of a polynomial p in $\mathbb{R}[x_1, \ldots, x_n]$ is homogeneous if and only if p is.

Visually: $p(x_1, \ldots, x_{n-1})$ defines a closed locus V within the hyperplane $x_n = 1$ of $(x_1, \ldots, x_n)$-space, and in this n-space, the locus of $H(p)$ is the closed homogenization of V. This forms the basic link between geometric and algebraic homogenization.

We can easily make this quantitative: the equation of the 2-subspace through any of the four lines L_n is $y = nz$. This describes the unique plane that passes through the origin and intersects the plane $z = 1$ in the line $y = n$. Of course, we actually started with four lines L_1, L_2, L_3, L_4 in the plane $z = 1$. Figures 2.3 and 2.4 could be fancied up to show the union of all four corresponding 2-subspaces instead of just one of them. Their union is still homogeneous.

The union and the intersection of any number of homogeneous sets is homogeneous.

In Figure 2.4 we'd get four heavy lines in the plane $x = 1$, all going through the point (1, 0, 0). These four lines are depicted in the next illustration, Figure 2.5. Figure 2.6 is an analogue of Figure 1.1 [p. 3], and summarizes what we've done.

Philosopher. I get the impression that you can interpret Figure 1.1 and Figure 2.6 both geometrically and algebraically. We are using it geometrically when we take an ordinary conic in $\mathbb{R}^2$, put it in the plane $z = 1$ and homogenize it by passing 1-spaces through its points. That's *geometrically* moving from "child" to parent. And slicing the cone with some other plane moves us back down to another child. In all this, the notion of "homogeneous" is geometric.

I'm wondering: is there some general *algebraic* procedure for going up one arrow and down another in Figure 1.1 or 2.6? For example, is there a way to go from an equation

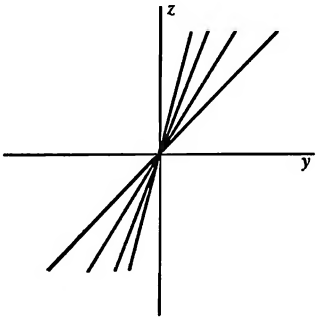


FIGURE 2.5. This picture answers the question raised in Figure 2.2. It shows what the four horizontal lines of that figure look like after we relocate to the "end" of any one of them.

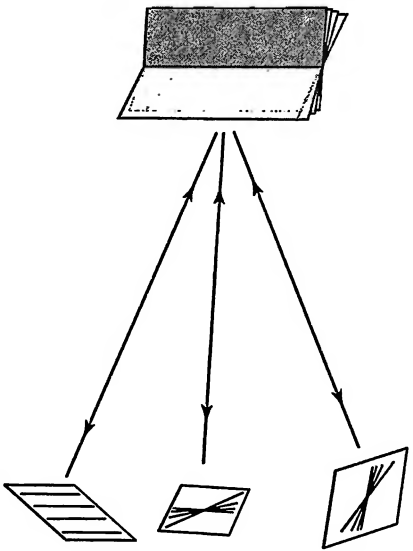


FIGURE 2.6. This is Figure 1.1 when the parent cone—a homogeneous set—happens to be the union of four planes through a common 1-space. The "child" at the lower left is Figure 2.2. The one at the lower right is Figure 2.5.

involving x and y, to the equation describing what we see when gazing down the x-axis, or down the y-axis?

Teacher. Yes, there is. To aid in explaining, let's introduce some definitions and notation.

Fact: There's a reverse notion to homogenization—it's called dehomogenization, and it also comes in both geometric and algebraic flavors. Here are the corresponding definitions:

Definition 2.2. Let C be a homogeneous set in $(x_1, \ldots, x_n)$-space. The intersection of C with the hyperplane $x_i = 1$ is called the **dehomogenization of C at x_i**, and is denoted by $D_{x_i}(C)$. Going from C to $D_{x_i}(C)$ is called **dehomogenizing C at x_i**. If C is a line, conic or curve, then $D_{x_i}(C)$ is called an **affine line, affine conic or affine curve.**

Definition 2.3. Let $p(x_1, \ldots, x_n)$ be a polynomial in $\mathbb{R}[x_1, \ldots, x_n]$. The polynomial obtained by setting $x_i = 1$ is called the **dehomogenization of p at x_i**, and is denoted by $D_{x_i}(p)$. Going from p to $D_{x_i}(p)$ is called **dehomogenizing p at x_i**. An **equation $p = 0$ is dehomogenized at x_i** by passing to the equation $D_{x_i}(p) = 0$.

One of the most powerful and frequently-used methods for investigating polynomially-defined curves (such as conics) is this: take an affine curve defined by a polynomial p, then homogenize p, and finally dehomogenize the result at another variable. Geometrically, we're "switching views."

This is important because...

the closed homogenization of any polynomial curve contains the maximum amount of information—an affine slice often contains less. On the other hand, it is in an affine view that we have available the full power of coordinates and algebra. Obligingly, every point of the homogenized curve, as well as its immediate environs, appears in at least one affine slice. Thus we get the best of two worlds—we can see every point, and we can also use coordinates and algebra to analyze what's happening around that point.

Fortunately, the promised procedure is both nice and neat, and works just as well for curves defined by a polynomial p. Here's what you do:

(1) First homogenize $p(x, y)$ to $H(p(x, y)) = P(x, y, z)$;

(2) Then dehomogenize with respect to whichever plane you want to look in. The polynomial in the plane $x = 1$, for example, is $P(1, y, z)$. In the plane $y = 1$, it is $P(x, 1, z)$.

These algebraically do exactly what you wondered about: in (1), p defines any one of the "children" at the bottom of Figure 1.1, and P gives the parent cone, whose equation is always homogeneous. Therefore (1) is an algebraic recipe for moving us *up* a line from child to parent. (2) is an algebraic recipe for moving *down* a line from parent to child.

Student. This algebraic recipe sounds powerful. Can you use it to show me what "the other end of a parabola" looks like?

Teacher. Yes, and that would nicely illustrate the power of this idea. Figure 1.13 [p.19] suggests that a parabola touches the line at infinity in one point, and you're wondering what you'd see looking directly down at that point at infinity. Well, inspired by this figure, let's consider the locus of $x + y^2 = 0$, which is a parabola opening to the left. The homogenization of $x + y^2$ is $xz + y^2$. As with the horizontal lines considered above, the infinite point of interest is at the end of the x-axis, so we use $x = 1$ as our screen. Setting $x = 1$ in $xz + y^2$ gives $z + y^2$. The parabolas $x + y^2 = 0$ and $z + y^2 = 0$ are essentially the same. This means that both "ends" of the parabola have exactly the same shape!

Philosopher. You seem to be implying that this method is not specific to lines or conics, but actually works more generally.

Teacher. Yes...

Student. This is just so incredible!

Teacher. It applies to any polynomial equation, and what it tells us is often quite enlightening. For example, many people know what the graph of the cubic $y = x^3$ looks like. But *what does it look like at infinity*? Certainly y grows rapidly compared to x, and we might suspect that the two "ends" of the graph approach a common point. If so, it must by symmetry be the point at the "end" of the y-axis. The homogenization of the polynomial equation is $yz^2 = x^3$. To view the graph from this new perspective, use the viewing plane $y = 1$. The dehomogenized equation is then $z^2 = x^3$. In Figure 2.7, the top two pictures show the usual plane and disk views of $y = x^3$. The bottom pictures are plane and disk views centered at the "end" of the y-axis:

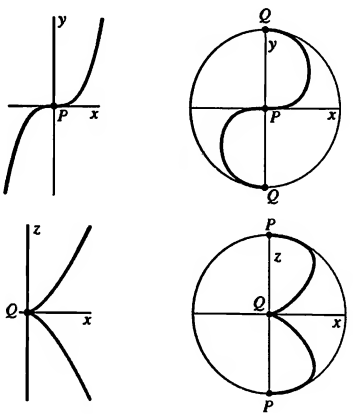


FIGURE 2.7. The top row shows the view of the cubic $y = z^3$, and its image in the disk model. The bottom row shows corresponding views after we relocate to Q, the point at infinity on the y-axis. The bottom row gives precise information on what happens at infinity: the ends of the cubic join together not smoothly, but in a "kink"—in the cusp curve $z^2 = x^3$.

Philosopher. I'd like to up the ante a bit! Suppose we consider a basic double cone intersecting the sphere in two opposite circles. The equation of one such double cone is

$$z^2 = x^2 + y^2. \tag{2.1}$$

Instead of rotating its intersection with a sphere, as we did in Figures 1.11–1.15, a space traveler could see the same thing by sweeping out an appropriate large circular orbit around the sphere. My question is this: can you apply your method to give me the equation of the double cone's intersection with an ordinary plane viewing screen kept perpendicular to the space traveler's rotating line of vision to the sphere? As we've done before, that screen can be taken tangent to the sphere. And for simplicity, take the sphere to have radius 1.

Teacher. Yes, we can do this. Look at Figure 2.8:

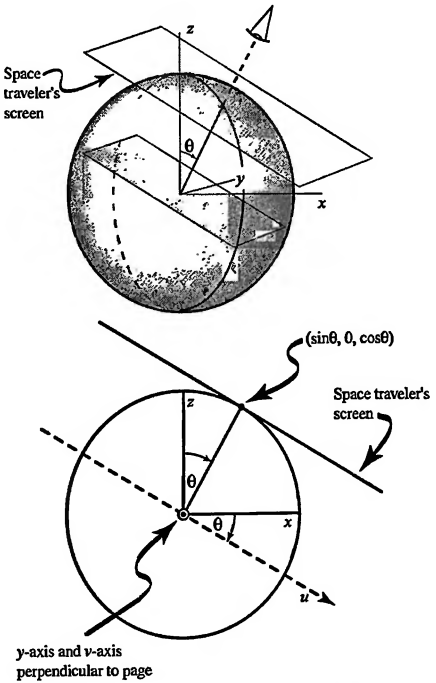


FIGURE 2.8. In the bottom picture, the y-axis is oriented directly away from us. The tangent plane and other planes are all seen edge-on, so they appear as lines.

Let θ, measured from the positive z-axis, be the angle swept out as the space traveler traverses a circle. The top picture shows the 3-D setup, with the space traveler's line of vision perpendicular to the viewing screen. The bottom picture is essentially the top one, but with the y-axis oriented perpendicular to the page.

In the bottom picture, we've taken a copy of the (x, y)-plane, calling its coordinates u (instead of x) and v (instead of y), and rotated this copy plane clockwise about the y-axis (that is to say, the v-axis) by an amount θ. From the figure we see that the (x, y, z)-coordinates of a point (u, v) in this rotated plane are $(u\cos\theta, v, -u\sin\theta)$. Next, translate this rotated plane by $(\sin\theta, 0, \cos\theta)$ to make it tangent to the sphere. This means that in Figure 2.8, we translate the plane through the origin to the tangent plane. The offset of $(\sin\theta, 0, \cos\theta)$ therefore leads to these parametric equations for the translated plane:

$$x = (\cos\theta)u + \sin\theta;$$

$$y = v;$$

$$z = -(\sin\theta)u + \cos\theta.$$

Substituting these into the cone's equation $z^2 = x^2 + y^2$ gives the restriction of the cone to our rotated plane. This simplifies nicely to:

$$(\cos 2\theta)u^2 + 2(\sin 2\theta)u + v^2 = \cos 2\theta. \tag{2.2}$$

Note that in (2.2) one can complete the square in u to get the standard form

$$\frac{(u + \tan 2\theta)^2}{(\sec 2\theta)^2} + \frac{v^2}{\left(\sqrt{\sec 2\theta}\right)^2} = 1. \tag{2.3}$$

For any θ, (2.3) descrived what the space traveler actually sees—the central projection of [sphere] $\cap$ [cone] onto the screen before his eyes.

> The verb *project* comes from the Latin *pro* ("forward") and *iactus* ("throw"). A projector literally throws film images forward onto the viewing screen; that's the underlying spirit of the term *projective plane*. Today, we understand the mathematics behind projecting—an understanding Leonardo da Vinci very much desired for drawing 3-D scenes on a flat surface.

Student. So a succession of conics given by (2.3), as θ increases, would amount to a "movie" of what the space traveler is seeing?

Teacher. Exactly. Choosing θ equal to multiples of $\frac{\pi}{8}$ gives flat-plane analogues of the pictures in Figures 1.11–1.15, producing "frames" of such a movie. Take a look:

$\theta = 0$: (2.2) reduces to the unit circle $u^2 + v^2 = 1$.

$\theta = \frac{\pi}{8}$: (2.2) can in this case be written as

$$\frac{(u+1)^2}{(\sqrt{2})^2} + \frac{v^2}{(\sqrt[4]{2})^2} = 1,$$

so the circle has become an ellipse, being stretched horizontally by the factor $\sqrt{2} \approx 1.414$ and vertically by the factor $\sqrt[4]{2} \approx 1.19$. The circle has gotten larger and, overall, stretched horizontally. The center of this ellipse is at $(-1, 0)$, but the original center point—the origin—has itself moved only to $(-\tan\frac{\pi}{8}, 0) \approx (-0.4142, 0)$. The right vertex of the ellipse seems to be approaching the origin.

Student. Why should the ellipse get *larger*?

Teacher. Equation (2.3) tells us that the center of symmetry shifts to the left by $\tan 2\theta$, that the semi-horizontal axis is $\sec 2\theta$, and that the semi-vertical axis is $\sqrt{\sec 2\theta}$. For

$$0 \le \theta < \frac{\pi}{4},$$

both semi-axes increase, so the ellipse grows. But the horizontal semi-axis increases faster than the vertical one, making the ellipse, overall, horizontally elongated.

A growing ellipse makes sense on purely geometric grounds, too. The cone with vertex at the center of the sphere can be looked at as a source of light rays, and the further away you get, the more separated the rays become. As we rotate the viewing plane through the angle $\frac{\pi}{8}$, it intersects most lines of the cone farther away, and these further-away intersections make the whole ellipse larger.

$\theta = \frac{\pi}{4}$: (2.2) now reduces to $2u + v^2 = 0$, a parabola opening to the left. The distortion that changed the original circle to a larger, horizontal ellipse, has truly reared its head: the expansion is *tremendous*—it's a figure that is unbounded, both horizontally and vertically. The horizontal expansion is relatively greater, producing, as a result, a leftward-opening parabola. The rightmost point of the figure, which was apparently heading toward the origin in the $\frac{\pi}{8}$ picture, has at $\theta = \frac{\pi}{4}$, actually reached it. The original center, (0,0), is at $(-\tan\frac{\pi}{4}, 0) = (-1, 0)$.

$\theta = \frac{3\pi}{8}$: For this θ, (2.3) becomes

$$\frac{(u+1)^2}{(\sqrt{2})^2} - \frac{v^2}{(\sqrt[4]{2})^2} = 1.$$

The above parabola's point at infinity has been pushed across the line at infinity, appearing now on the right side. The central point between the two branches is $(-1, 0)$, and the original center, $(0, 0)$, has shifted to $(-\tan\frac{3\pi}{8}, 0) \approx (-2.4142, 0)$.

$\theta = \frac{\pi}{2}$: (2.2) reduces to $u^2 - v^2 = 1$, a hyperbola symmetric about the origin. The original origin is now at $(\pm\infty, 0)$.

With further $\frac{\pi}{8}$ increments, the hyperbola continues moving left. When $\theta = \frac{3\pi}{4}$ it has morphed into the rightward-pointing parabola $-u + v^2 = 0$. Finally, $\theta = \pi$ completes the cycle, giving the original circle $u^2 + v^2 = 1$. Algebraically, the period of $\theta = \pi$ makes sense, since in (2.2), θ always has a "2" as its coefficient.

> In the family (2.2), the parabola corresponding to $\theta = \frac{\pi}{4}$ is a bridge between the ellipses and hyperbolas.

There always must be some sort of bridge when you morph between ellipses and hyperbolas, but that bridge need not be a parabola. For example, in the family $kx^2 + y^2 = k$, we get an ellipse when $k > 0$ and a hyperbola when $k < 0$; the transition curve at $k = 0$ is a double line. In the family $x^2 + 2kxy + y^2 = 1$, we have an ellipse when $|k| < 1$ and a hyperbola when $|k| > 1$; the bridge corresponds to $k = \pm 1$ and consists of two parallel lines.

An Unexpected Bonus

Teacher. In changing views, we've thought of $p(x, y)$ as defining a curve in the viewing plane $z = 1$. If this curve happens to be given parametrically, say as $(x(t), y(t), 1)$, one can change the view to the plane $x = 1$ (or to the plane $y = 1$) without homogenizing. Geometrically, dividing each component of $(x(t), y(t), 1)$ by a common nonzero real number just slides the point along a line through the origin. That is what's happening when we change views. If we look at a point in [cone] $\cap$ [plane] while slowly rotating the plane, we see the point of intersection sliding along a particular line through the origin. Suppose that after dividing through by some such number, the first component happens to become 1. Then we have a view in the $x = 1$ plane, and of course this can be attained simply by dividing through by $x(t)$, as long as $x(t) \neq 0$. This gives a new parametrization

$$\left(1, \frac{y(t)}{x(t)}, \frac{1}{x(t)}\right).$$

Similarly, dividing through by $y(t)$ gives

$$\left(\frac{x(t)}{y(t)}, 1, \frac{1}{y(t)}\right),$$

a parametrization of the cone's intersection with the plane $y = 1$.

We can apply this idea to the circle in the plane $z = 1$ parametrized by $(\cos t, \sin t, 1)$. The cone this circle defines is $z^2 = x^2 + y^2$, and our method says that a parametrization of this cone's intersection with the plane $x = 1$ is

$$\left(1, \frac{\sin t}{\cos t}, \frac{1}{\cos t}\right) = (1, \tan t, \sec t).$$

By suppressing the first coordinate, we can identify the intersection with a hyperbola in the (y, z)-plane, with branches opening in the z-direction, and we have thus found for it a new parametrization, $(\tan t, \sec t)$. Similarly, the intersection with the plane $y = 1$ plane gives a hyperbola in the (x, z)-plane, with branches opening in the x-direction, and leads to the parametrization $(\sec t, \tan t)$.

Philosopher. These seem simpler than the more familiar parametrizations using $\cosh t$ and $\sinh t$.

Teacher. I agree. The familiar hyperbolic parametrization $(\cosh t, \sinh t)$ doesn't trace out the entire hyperbola as t runs through all real values—only the right branch. One has to

use an additional parametrization $(-\cosh t, \sinh t)$ for the left one. But the corresponding circular-function parametrization $(\sec t, \tan t)$ *sweeps out the entire hyperbola as* t *advances by* 2π. This brings out the essential loop nature of the hyperbola, helping to unify it with a circle or ellipse. The parametrization even tries its best to produce the two points at infinity, where this hyperbola intersects the line at infinity. This happens at $t = \frac{\pi}{2}$ and $t = \frac{3\pi}{2}$, where both sec and tan become infinite.

> The circular-function parametrization $(\sec t, \tan t)$ sweeps out an entire hyperbola as t advances by 2π. The more familiar hyperbolic parametrization $(\cosh t, \sinh t)$ gives only one branch as t runs through all real numbers. Using circular functions brings out the loop nature of the hyperbola.

Why Is Mathematics Like the Hawaiian Islands?

Philosopher. I have to say it looks like we've succeeded in bringing out some of the unity among conics, as I'd hoped. It brings to mind my thoughts of why Mathematics is like the Hawaiian Islands.

Student. A story? Go for it!

Philosopher. Hawaii may *look* like a group of separate islands, but actually the islands are just the highest peaks of an immense, mostly-submerged mountain range. All that water hides their underlying connectedness, their oneness. Mathematics may similarly seem like an archipelago of different areas—geometry, analysis, topology, number theory, applied math, and so on. My philosophy is that we're really just seeing a few peaks of a huge mathematical mountain range. Our ignorance is like the water surrounding Hawaii and hiding its true mountain-rangeness. In mathematics, when we remove ignorance by making discoveries and advances, the water level in effect goes down, and when it drops far enough, separate islands get connected.

A good example of two islands joining is the discovery by René Descartes (1596–1650) of analytic geometry. Before that, algebra and geometry led mostly separate lives. The ancients, for example, had no equation such as $x^2 + y^2 = 1$ to describe a circle. Proofs of geometric facts took place within the geometric realm, and as a consequence, finding a proof often depended on inspiration or cleverness. Today, we can simply translate geometric assertions into algebra: points, lines and circles become point-coordinates, line-equations and circle-equations. The geometric punch line of an assertion becomes a corresponding algebraic punch line, usually involving equations of degree 2 or less. The task is then one of establishing the equation's truth. Though this can involve more steps, it is usually more straightforward.

Teacher. That's a vivid example. It's interesting to point out that the symbiosis between algebra and geometry also has a more modern, abstract counterpart. In algebraic topology, algebraic tools such as groups and homological algebra shed light on various ways of classifying topological manifolds, for example.

It's not only mathematics that is like the Hawaiian Islands—physicists also are engaged in a constant struggle to achieve unity. Today's biggest problem: unifying quantum mechanics with the general theory of relativity.

Albert Einstein (1879–1955)

Einstein's whole life was a monument to unification. In his special theory, mass and energy become "mass-energy" through $E = mc^2$. Also, space and time—separate in Newton's theory—become space-time. Then in his general theory he showed how mass-energy and space-time interact: mass-energy tells space-time how to curve, and the curving tells mass-energy how to move.

Chapter Summary

In Chapter 1, we saw that a rotating sphere with two antipodal ellipses drawn on it forms a visual bridge between ellipses and hyperbolas, even uniting asymptotes with certain tangent lines. This chapter adds an algebraic component, making Chapter 1's qualitative results quantitative. The technique of "homogenizing/dehomogenizing" the polynomials defining lines or conics gives the new views concrete equations, allowing one to explore things in greater detail. As an ellipse picture morphs into a picture of a hyperbola, so do their equations, lending greater algebraic unity to our study.

Exercises

1. (a) Prove that the locus in $\mathbb{R}^3$ of a polynomial equation $p(x, y, z) = 0$ is homogeneous if and only if p is homogeneous.
   (b) Notice that $x^2 + y^2 + z^2 = 0$ defines a singie point in $\mathbb{R}^3$. That isn't a problem in $\mathbb{P}^2(\mathbb{R})$, is it?
2. (a) Assume the four heavily-drawn lines in Figure 2.2 on p. 31 have equations $y = n$ $(n = 1, 2, 3, 4)$. Show that the equation of their union is
$$(y - 1)(y - 2)(y - 3)(y - 4) = 0.$$
   (b) Using the homogenization/dehomogenization technique described in the text, translate the pictures in Figure 2.2 and Figure 2.5 (on pp. 31 and 34) into algebraic terms, thereby obtalning an analogue to (a)—one equation for the union of the four lines in Figure 2.5.
3. Write parametric equations for the cubic and cusp curves appearing on the left side of Figure 2.7 (p. 36), and substitute them into the plane-to-disk mapping in (1.1) (p. 12) to get parametric equations for the right-hand parts of Figure 2.7. Have Maple plot these

last equations. (You now have a technique for plotting all sorts of curves in the disk view!)

4. Draw curves analogous to those in Figure 2.7 for the hyperbola $xy = 1$.

5. Draw curves analogous to those in Figure 2.7 for the curve $y = x^4$.

6. The homogenization of the cubic $y = x^3$ is, of course, $yz^2 = x^3$. Have Maple draw this surface within the unit cube having faces $|x| = 1$, $|y| = 1$, $|z| = 1$. Now appropriately rotate your figure so you can see the edge of the surface on the face $z = 1$. That edge is the cubic $y = x^3$, so you're viewing the dehomogenization of the homogeneous surface at $z = 1$. Similarly view its dehomogenizations at $x = 1$ and $y = 1$.

7. What degenerate conic serves as a bridge connecting different parts of the family of hyperbolas $xy = k$ ($k \in \mathbb{R}$)?

8. There's a "soul mate" to the usual hyperbolic parametrization of a branch of the hyperbola $x^2 - y^2 = 1$: a hyperbolic parametrization for a semicircular "branch" of the circle $x^2 + y^2 = 1$. Find a parametrization for each such branch. As a check, have Maple sketch your branches.

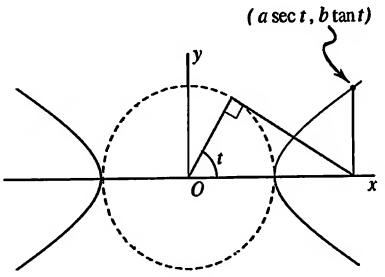
9. (a) By using the method introduced in this chapter, generalize the parametrization $x = \sec t$, $y = \tan t$ for the hyperbola $x^2 - y^2 = 1$, to $x = a \sec t$, $y = b \tan t$ for

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1.$$

(b) Assume you know nothing about homogenizing/dehomogenizing. Verify the coordinates shown below for an arbitrary point on the hyperbola

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1.$$

The circle has radius a.



3

How to Gift-Wrap a Hyperbola

in which another exception falls:
like ellipses, every hyperbola
and parabola lives in a box

Do Hyperbolas Live in Boxes?

Philosopher. I've been thinking about this picture, showing three conics as they might appear in a textbook:

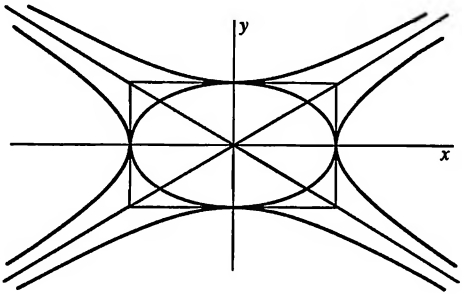


FIGURE 3.1. Three conics in standard form.

It seems to me that the rectangle ought to play the same role for hyperbolas as for ellipses. Certainly, in Figure 3.1, it doesn't.

Teacher. Explain.

Philosopher. The rectangle is useful in drawing both the ellipse and the hyperbolas, but in different ways. The ellipse is tangent to the rectangle at the midpoint of each of its four sides, and one can use this to fill in the whole ellipse. But for either hyperbola, the rectangle's diagonals become important, giving us the asymptotes. They're not used at all in drawing the ellipse!

Teacher. I found myself in a similar predicament when you asked "Why don't ellipses have asymptotes?" Your question seemed insoluble—even somewhat nonsensical—until we morphed between ellipse and hyperbola. Then we saw the truth: an asymptote is nothing more than a tangent line, except that the point of tangency happens to be at infinity. My gut instinct is: if an idea works once, try it again!

Philosopher. Morph again?

Teacher. Yes. Starting with an ellipse and its surrounding rectangle, morph the ellipse into a hyperbola. Follow the box along, and see what happens.

Student. That sounds clever!

Philosopher. So we rotate a sphere, as before?

Teacher. Yes, we can rotate by increments, producing a series of pictures. I think it would be helpful to look at both the disk picture and corresponding plane picture. Figures 3.2–3.5 show such a sequence, in which we've divided the full 2π rotation into 16 roughly equal increments. Working our way through them may shed some light on the question.

What Do All These Pictures Tell Us?

Student. The top row of Figure 3.2 is pretty obvious, isn't it?

Teacher. Yes. A circular cone defines a circle in this initial view. In each picture of this pair, the box is heavily drawn. I've also numbered the sides, and put arrows on them. We'll track everything mercilessly!

Philosopher. Looks like we should learn *something*!

Teacher. We now come to the second row in Figure 3.2. Notice, in the affine view, how the rectangle has changed into a quadrilateral.

Philosopher. I don't like that. The thing that's supposed to surround an ellipse is a *rectangle*, not some other quadrilateral.

Teacher. But this is what we get when we orthogonally project the sphere image onto the viewing plane, as we rotate the sphere!

Philosopher. Then maybe rotating the sphere is not the way to go.

Teacher. But it has already answered some of your questions. Let's not throw away something that's already worked. Actually, far from being arbitrary, that quadrilateral is very informative.

Philosopher. In what way?

Teacher. Remember how basic the idea of distortion is in conic sections? We've started with a circular cone. In the top row of Figure 3.2, no rotation has taken place, so we get a circular image. In the second row, right, we see that the rotation has distorted the circle into an ellipse, and the square, into a quadrilateral. In fact, that quadrilateral almost screams "distortion"! Its obvious non-rectangularity provides more specific information than the ellipse does.

Philosopher. Explain.

Teacher. Consider: one can use a square and its inscribed circle to define a circular cone inside a right pyramid with square base. A plane not through the vertex cuts the cone in

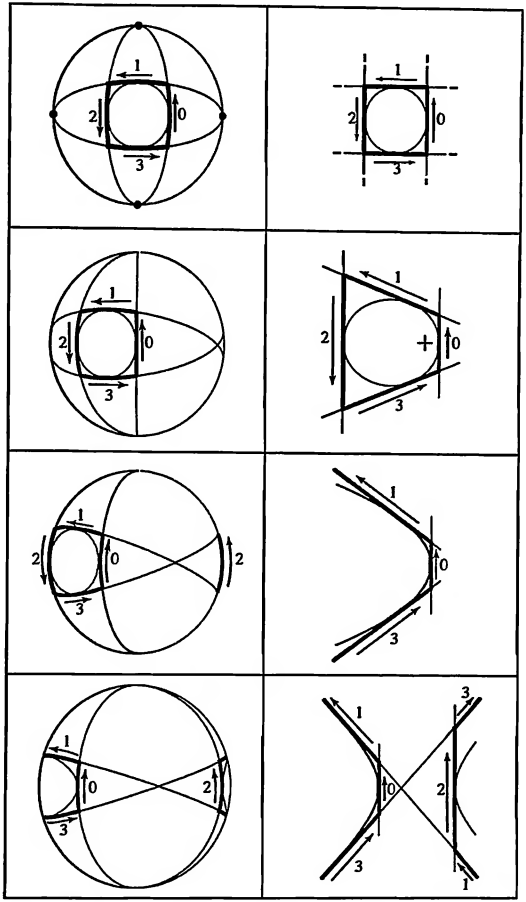


FIGURE 3.2. Here are the first four of 16 pairs of views tracing the history of the standard rectangle surrounding an ellipse as we morph it. For clearer drawings, we use the circle $x^2+y^2=r^2$, with $r \approx 0.4$. The top row shows the disk and affine views of a front-and-center shot of the circle. In the second row, right, we see in the affine plane a circle elongated horizontally. By the third row it has become so elongated that it's now a parabola opening to the left. In the bottom row, left, part of the circle has moved across the line at infinity, and in the right view we see it as a hyperbola.

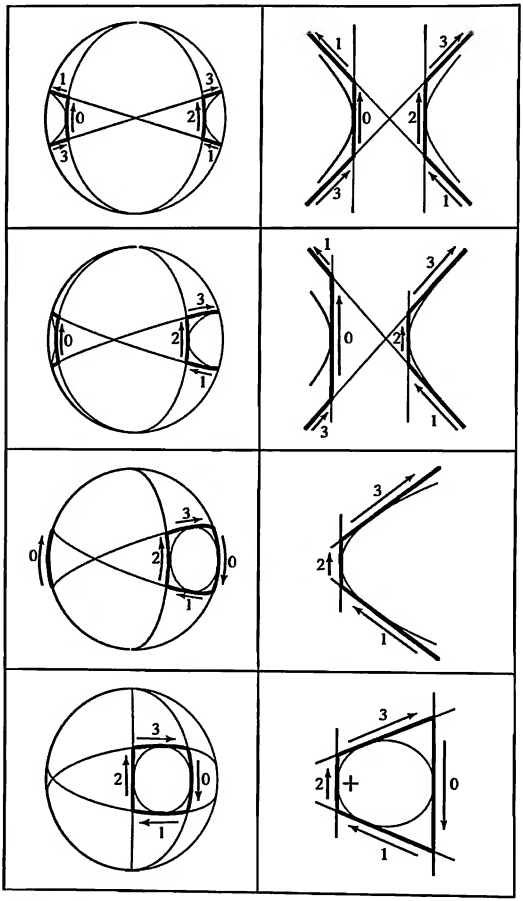


FIGURE 3.3. In the top row, we see the original circle after a rotation by $\pi/2$; on the right it appears as a front-and-center view of the hyperbola $x^2-y^2=r^2$. By the third row, the rightmost point of the original circle has moved to infinity, and we again see a parabola, this time opening to the right. In the bottom row, left, the parabola has moved away from the line at infinity, and we see an unbroken loop, which in the affine plane is a horizontally-elongated ellipse. It's translated as much to the right as it was to the left in the second row of Figure 3.2.

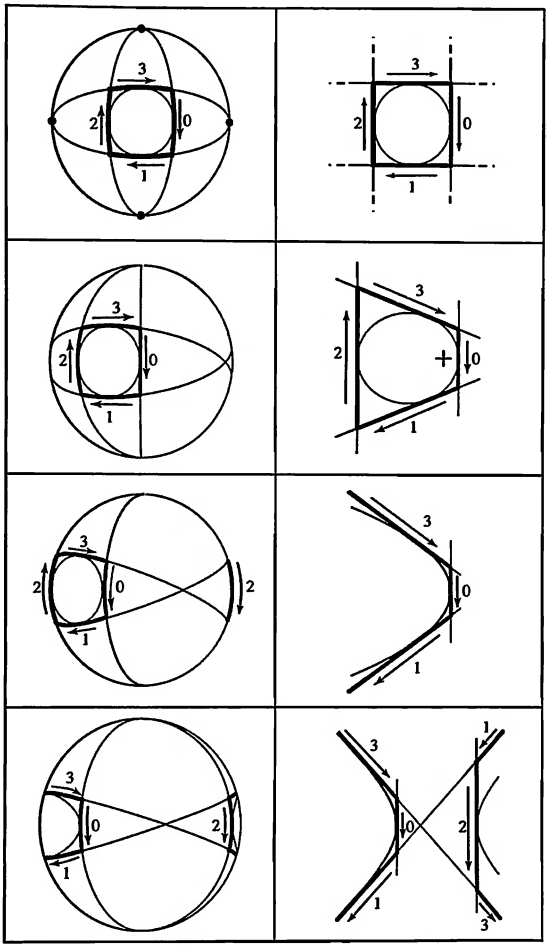


FIGURE 3.4. The top row corresponds to rotation by π, and has similarities with the top row of Figure 3.2. Already by the third and fourth rows of Figure 3.2, however, we see one edge reversing its orientation as it passes across the line at infinity. A basic difference between this figure and Figure 3.2 is that, after a rotation by π, every edge here has crossed infinity once, and the orientation of the surrounding square is reversed.

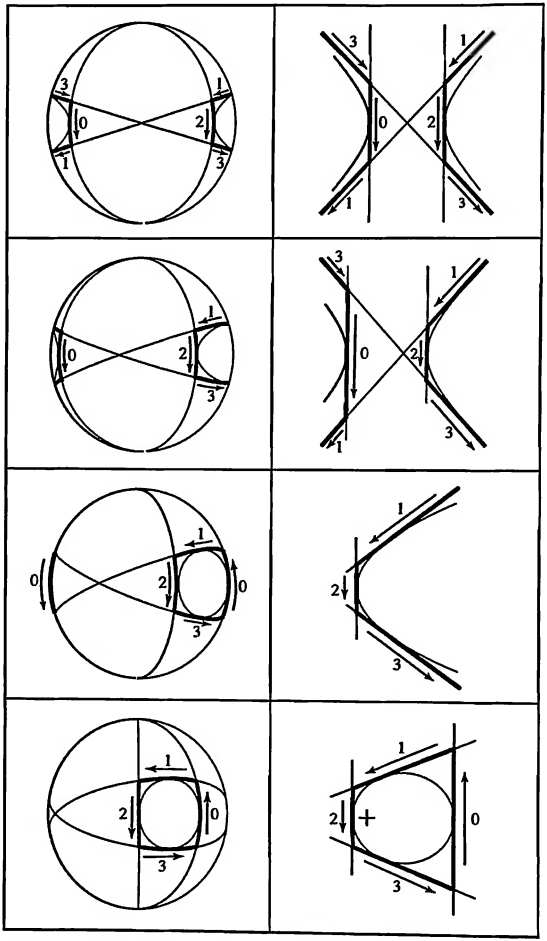


FIGURE 3.5. This figure is similar to Figure 3.3, but box orientations are reversed. After the bottom row, an additional rotation brings us to a full 2π, and back to the top row of Figure 3.2.

an ellipse and cuts the pyramid in a quadrilateral which in general isn't a square, or even a rectangle. The "non-squareness" of the quadrilateral communicates and measures the actual distortion better than the ellipse does.

Philosopher. But the ellipse is so symmetric!

Teacher. That pleasantness is also the problem. Fact is, an ellipse has a certain amount of "magic" to it—although that plane intersects the pyramid to create a quadrilateral with *some* symmetry, it cuts the cone in a figure that has decidedly greater symmetry, an ellipse. But this very magic also masks the full extent of the distortion!

> "Ellipse magic" has yet another face: you can try to shear an ellipse to force its principal axes away from mutual perpendicularity. To no avail. In shearing, the images of the principal axes don't stay principal; instead, new directions become principal. The ellipse always wins: principal axes remain mutually orthogonal.

Philosopher. I still have a problem with these quadrilaterals: although they're central projections of the sphere's great circles, or "lines," we've lost the power that a rectangle has. Given a rectangle, there's exactly one ellipse tangent at the midpoint of each side—the rectangle completely determines the ellipse. You don't generally have that with a quadrilateral.

Teacher. But there's an analogue. Notice that our quadrilateral has one axis of symmetry. It turns out that *there's exactly one ellipse sharing the quadrilateral's axis of symmetry and tangent to each side of the quadrilateral.*

Philosopher. Even so, with the rectangle, the uniqueness goes the other way, too: given an ellipse, there's a unique box surrounding it. Do you have *that* for quadrilaterals?

Teacher. No, but the lack of uniqueness more fairly reflects the actual range of possibilities. For example, if the plane intersects an elliptical cone orthogonally, we get a front-and-center mug shot, and what surrounds the ellipse is a rectangle. But not all photos are front-and-center! We could get that same ellipse another way: start, say, with a *circular* cone and a plane which intersects the cone's axis *non*-orthogonally. Now the quadrilateral surrounding that ellipse isn't a rectangle. From this quadrilateral, one can actually reconstruct the angle the plane makes with the circular cone's axis. The angle introduces distortion, and the quadrilateral faithfully records that distortion. Furthermore, one can even reconstruct the cone itself. The rectangle doesn't seem even to acknowledge the possible variations in angle.

Philosopher. Well, you make a persuasive case. I want my question answered, so let's proceed!

Teacher. In going from row 2 to row 3 in Figure 3.2, a lot happens. As side 2 rotates toward the horizon, it encounters the tremendously-increasing concentration of metric there. When we actually arrive at the third row of Figure 3.2, side 2 has undergone truly severe distortion. It has moved onto the horizon, and therefore no longer has finite length. Of course one end of side 1 is also on the horizon. the other end remaining in the vicinity of

center field (the same is true for side 3), so these lengths also are no longer finite. The disk view shows all four points of tangency with the box, and the affine view shows the three of these that are in the finite plane. Notice that side 0 has shrunk to the shortest it will ever be, since it passes through center field. As we rotate further, it will begin to stretch.

Philosopher. That affine view in the third row of Figure 3.2 is remarkable—I never suspected one could draw a box around a parabola!

Teacher. Three sides, anyway, with clear points of tangency.

Philosopher. Nonetheless, let me play the devil's advocate, here. In the affine view, suppose at each stage one draws a nice conventional rectangle around the growing ellipse. What do you see in the limit?

Teacher. Go back to the standard form for the affine view of the unit circle, in (2.3) [p. 38]. The ellipse's semi-major axis is $\sec 2\theta$, and the semi-minor axis is $\sqrt{\sec 2\theta}$. Both of these grow without bound as θ approaches $\frac{\pi}{4}$, but of course the major axis length far outpaces the minor axis length in this. If we continuously back away from the plane to keep a total view of the ellipse in our field of vision, we see an ellipse getting ever thinner in the horizontal direction. If we back away in such a way that the changing ellipse always has a *constant* horizontal diameter, then in the limit the ellipse just collapses to a double line segment—two segments positioned in the same space. This limit is not at all as informative as the affine view in the third row in Figure 3.2.

Philosopher. But suppose you don't back away so fast—say only fast enough to keep the increasing *vertical* side of fixed length?

Teacher. Then the horizontal sides become unbounded, and the point of tangency rides along to infinity. In the limit, the ellipse degenerates to a double ray pointing leftward. Again, this doesn't provide as much information as in the affine view.

Philosopher. You do persuade.

Student. I'm noticing something strange here. In the disk view of the third row of Figure 3.2, you have two side 2's.

Teacher. Actually, only one, because antipodal points are identified. They're the same side.

Student. Yes, but you've marked them with arrows going in opposite directions!

Teacher. That's just a consequence of identified points being antipodal. If you travel down the left copy of side 2, then its antipodal image goes *up* on the right. Your observation is equally noteworthy in the fourth row. In the disk view, sides 0 and 2 have arrows going the same way, while those on sides 1 and 3 don't.

Philosopher. As soon as we rotate beyond a certain point, terminology suddenly changes, and we call it a *hyperbola*. That bothers me!

Teacher. I understand, and one starts talking about branches, too. Such arbitrary terminology does a real job in camouflaging the underlying unity! It reminds me of your "Hawaiian Islands" analogy.

Philosopher. The fact that we don't see the line at infinity in the affine model is like the "ignorance is water" part of my analogy, clouding the unity that's really there.

Teacher. Yes. However, despite all the drawbacks, it's admittedly convenient to refer to left and right branches.

Philosopher. Unfortunately, it looks like you've made a little mistake in drawing the bottom row of Figure 3.2. In the affine view, I see the asymptotes for the left branch, but you've got the right branch too far to the right. What happened?

Teacher. The heavy lines used to track the quadrilateral have obscured something important. In the third row of Figure 3.2, we see that sides 1 and 3 are each tangent to the parabola. With just the slightest additional rotation of the sphere, each of these tangent points moves just a bit farther away from center field, but nonetheless they're still points of tangency. The two branches do form a hyperbola, and there are asymptotes to it—but our lines are *not* those asymptotes!

Philosopher. What are they, then?

Teacher. They're the lines determining the quadrilateral enclosing the hyperbola. They're tangent to the hyperbola, but not at infinity, so we don't call them asymptotes. Of course, the quadrilateral now superficially looks broken, but is connected since opposite boundary points are identified. We see this in the disk view.

Philosopher. *So it appears, at last, that I see a box surrounding a hyperbola*! I'm no less impressed than by the one surrounding a parabola.

Teacher. The difference between our lines, and the asymptotes to the hyperbola, is essentially the same difference that we met between a rectangle surrounding an ellipse, and a more general quadrilateral surrounding it. We see this, for example, in the affine view on the bottom row of Figure 3.2. Here, we see differing lengths of sides 0 and 2. One can think of side 2 as coming fresh from infinity, and therefore longer than side 0, which is only now starting to head there. We're not yet looking at a front-and-center photo of a hyperbola.

Philosopher. But at the next stage, the top row of Figure 3.3, we are?

Teacher. Yes. This is similar to a rectangle surrounding an ellipse. Notice that in that top row, the four points of tangency are symmetric with respect to the origin, just as they are in the ellipse/rectangle case.

Philosopher. I guess we could push a point a little, and say that just as the rectangle surrounding an ellipse has opposite sides of equal length, the same is true here: sides 0 and 2 in the affine view on this top row have the same finite length, while sides 1 and 3 have infinite length.

Teacher. The symmetry in this figure does seem to make your intuitive statement a little safer-sounding.

Philosopher. I see that by the second row of Figure 3.3, the sphere has turned beyond $\pi/2$, and we've again lost symmetry, just as in the second row of Figure 3.2.

Teacher. Yes, and in the third row of Figure 3.3, the sphere's circle is on the right side of the disk, and we see the mirror-image parabola in the affine view. In the bottom row there, we see a mirror-image version of the quadrilateral in Figure 3.2's second row, and finally, the original circle in the top row of Figure 3.4.

Student. But that's different from the top row of Figure 3.2—all the arrows are going the other way! What's going on?

Teacher. Good point. Actually, something very fundamental is going on. An oriented loop, such as the original square, reverses its orientation when pushed past the line at infinity. We can compare this with a Möbius strip constructed from a long thin rectangle by first

giving one end a 180° twist, then taping the ends together. An oriented loop, pushed all the way around the Möbius strip, comes back to its starting position with its sense reversed. If this reversal can happen on a surface without crossing boundary points, we call the surface **nonorientable**. Indeed, such a reversal shows that it's impossible to assign an orientation at each point in a consistent way. If no such reversal is possible, the surface is called **orientable**. That 180° twist in the Möbius strip is actually more than merely reminiscent of your previous observation about how the arrow of side 2 changed direction when its opposite image appeared on the right side of the disk. There's a close relation between our real projective plane, and the Möbius strip.

Student. Really? I'm curious!

Teacher. Here's a little experiment: on the disk model of the projective plane, draw with an ordinary ruler two horizontal lines equidistant from the disk's center. The portion of the disk between these two horizontal lines defines a strip, and identifying antipodal boundary points of this strip is topologically the same as giving a 180° twist and taping the edges together.

Student. A Möbius strip!

Teacher. Precisely. The real projective plane simply *continues the taping process used in making the Möbius strip.*

Student. What do you mean?

Teacher. Well, visualize a little ant sewing together two opposite rectangle sides in the usual way of making a Möbius strip; when those two sides are entirely attached, he's sitting where two rectangle corners meet, and he's got a completed Möbius strip. I'm proposing that he *continue* his sewing. He's actually sitting on a boundary point, a boundary that hasn't been sewn together yet, and he sees some of this boundary to his right, and some to his left. So have him just start zipping these together, and let him sew this way until he has stitched together *all available boundary points.* We can see this continuation idea expressed in the arrows in two of the pictures in Figure 1.7 [p. 10]: the "before commencing the continuation" is the Möbius strip, and the "after he's all done" is the real projective plane.

Student. Let's make a model!

Teacher. I think I made that sound too easy. Unfortunately, the going gets tougher as we continue taping, and it's impossible to complete the task in our 3-dimensional world without self-intersection. That's why the disk or sphere model is so convenient: we can use familiar objects that are easy to visualize, together with simple taping instructions for them which can be performed in our minds, though not in 3-dimensional space. Actually carrying out those instructions? Often that can be done only in $\mathbb{R}^4$.

Philosopher. It seems that we're talking mathematics versus Mathematics, again. Our identifications are concessions to our human limitations—most of us can't visualize four-space, so we try various methods to help us out. Our valiant efforts are part of mathematics. I'd hazard a guess that "the real thing," the projective plane, smooth and sitting comfortably in four or more dimensions, is part of Mathematics.

Teacher. An interesting topic for debate. Anyway, our first eight pairs in Figure 3.2 and Figure 3.3 form half the total cycle, and the next set of eight in Figure 3.4 and Figure 3.5,

complete the full cycle. A picture after the last row of Figure 3.5 would coincide exactly with the first row of Figure 3.2, with all arrow directions agreeing. One can use the whole sequence to follow along what happens to the quadrilaterals through the entire cycle.

Student. Could we say that the whole sequence is like going around a Möbius strip *twice*, so that all original orientations are unchanged?

Teacher. Yes.

Philosopher. What about my original picture in Figure 3.1, and my question—the one that inspired this sequence of pictures?

Teacher. We can answer your question this way: First, I agree that the single rectangle in your picture, Figure 3.1, does function differently for ellipses than for hyperbolas. The unity you are seeking comes from realizing that there's not just one box that functions for them all, but that ***each of those two hyperbolas, as well as an ellipse, has its own box***. However, *each of these boxes functions in an identical way*. In fact, for every angle of sphere rotation—for every view—there's a box, or quadrilateral, within which one can draw the conic, even if it's a parabola. In your Figure 3.1, in addition to the rectangle, I could draw two other boxes, derived from the top, right picture of Figure 3.3. The first new box would consist of sides 0, 1, 2, 3 from that picture, with the finite sides vertical. The other new box would look roughly like a 90° rotation of the first box, its finite sides being horizontal. I'd therefore *delete the finite diagonals of the rectangle in Figure 3.1, leaving everything else about your picture the same*. The finite box is then surrounded by two other boxes. Each box contains a conic that touches the box tangentially at four points symmetrically situated with respect to the origin.

Philosopher. That does help clarify things. We're making progress!

Chapter Summary

In this chapter we see that the fundamental rectangle surrounding an ellipse has a hyperbola-analogue—any hyperbola has a "box" surrounding it. In a certain sense, that's true of a parabola, also. We present a sequence of pictures translating between a rotating sphere and the customary real-plane view, helping to concretize the evolving unity of our study.

Exercises

1. Draw the images in the disk model of the three circles $x^2 + y^2 = r^2$ in $\mathbb{R}^2$, where $r = \frac{1}{2}, 1, 2$.
2. A square pyramid has vertices (1,1,0), (−1, 1, 0), (−1, −1, 0), (1, −1, 0), (0,0,1) in (x, y, z)-space. Suppose that a plane containing one of the edges in the (x, y)-plane cuts the pyramid in a quadrilateral, and let α be the usual (dihedral) angle this plane makes with the (x, y)-plane. Find, in terms of α, the side lengths of this quadrilateral.
3. In Exercise 2, inscribe a circle in the pyramid's square base and let C be the cone containing this circle with vertex (0,0,1). The plane in Exercise 2 then intersects the cone in an ellipse.

(a) In terms of α, find the ratio of the ellipse's major axis length to its minor axis length.

(b) For each side of the quadrilateral, find the point of tangency with the ellipse.

4. The third row, right picture in Figure 3.2 shows three sides of the bounding box of a parabola. If the circle used has radius $r = 1$, then where, exactly, do the upper and lower sides touch the parabola? (Equation in (2.2) on p. 38 may be helpful in solving this problem.)

5. Show that $\mathbb{P}^2(\mathbb{R})$ can be constructed by sewing together a disk and a Möbius strip along their boundaries.

4

The Cube

in which equality among conics is put into an especially symmetric form.

Paradise Lost?

Philosopher. We've just found that the following picture solves the "box problem" arising from Figure 3.1 [p. 45]:

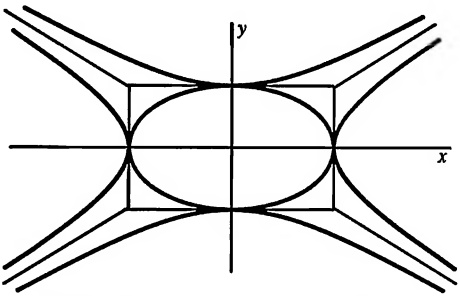


FIGURE 4.1. To construct the traditional rectangle surrounding an ellipse, one removes excess parts of two horizontal and two vertical lines. But there are two more lines: the asymptotes. Regarding the rectangle's two diagonals as their "excess"—and removing that excess—completes the last third of a basic process. The final result is that each conic is surrounded by a "quadrilateral."

Yet, for all the unity we're hoping for—for all the equality between hyperbolas and ellipses we've been promising—in Figure 4.1 the ellipse and hyperbola still look so *different* from each other! We surely can't expect this image to be the final, victorious answer to our quest for equality.

Teacher. A plane view introduces distortion, and in Figure 4.1 that distortion ruins any chance for equal-looking conics. Of course dealing with distortion is the main reason we

introduced the sphere. If we translate Figure 4.1 to a sphere with opposite points identified, we get three conics.

Philosopher. Identifying points means we'd actually be looking at *six* loops on a sphere.

Teacher. That's right. Notice that the ellipse in Figure 4.1 touches each of the other two conics in two points. On the sphere model of $\mathbb{P}^2(\mathbb{R})$, this extends nicely: *any one of the three conics tangentially touches the other two in two places.*

Philosopher. Yes—it's because points at infinity are no longer missing. But there is also the matter of the *lines* in Figure 4.1. They translate to great circles on the sphere.

Student. We've barely started talking about this problem, and already things are getting crazy. On a sphere, you have six smaller loops as well as those great circles, all trying to live with each other on a sphere—and you're looking for the *ultimate picture*?! This doesn't sound like a good idea to me.

Philosopher. I have to agree with you. Visualizing all this on a sphere *is* hard. We made a big push for doing things on a sphere, but distortion or not, visualizing things on a flat plane is easier. I wonder if there's some alternative way that has the advantages of both plane and sphere—some other screen on which to project light rays.

Teacher. If we're seriously going to look for another screen, we should keep in mind that the one thing remaining constant is the point light-source itself, which in this case creates three cones tangent in pairs.

Philosopher. One elliptical and two hyperbolic cones...

Teacher. Yes, but the hyperbolic cones are hyperbolic only relative to the plane screen $z = 1$. Relative to either $x = 1$ or $y = 1$, each of them is also elliptic and thus on a more equal footing.

Philosopher. But $x = 1$, $y = 1$, $z = 1$ are mutually orthogonal planes. Are you looking for a screen that uses all three planes *simultaneously*?

An Inspiration

Student. What about the walls of a room? A few years ago I made a sort of planetarium. For a "dome" I used an ordinary room.

Teacher. Maybe you've hit upon an idea! Perhaps one can think of a cube as occupying a sort of middle ground between a plane and a sphere.

Philosopher. And I thought *I* was far out!

Teacher. Seriously, it seems to me that a sphere acts a lot like a many-faceted polytope in which each small face intersects the cone orthogonally, or very nearly. We thus get limited distortion within each face. At the other end of the spectrum is a "polytope" having just one, huge face—a plane!

Philosopher. I think I see your basic idea. The polytope should be centrally symmetric, and opposite faces identified. You could think of two parallel planes, with the origin midway between, and identifying opposite points. If you include the planes' line of intersection at infinity, you even have an edge, to make it a "two-tope."

Teacher. Yes. But for our purposes a cube is better than a plane, even though it's still a "poor-man's sphere," with only six faces. Fact is, the finite extension of each face limits

the amount of distortion, and we can easily visualize what is on each face because it's flat.

Philosopher. With those three elliptical cones tangent in pairs, it'd be important to have one ellipse on each face, wouldn't it?

Teacher. I agree. We should avoid having any ellipses spilling over onto other faces. We could use a rectangular box instead of a cube; that would give us additional freedom. In dehomogenizing, we can set x equal to any nonzero constant; likewise for y and z.

Philosopher. We can't forget that there are those lines in Figure 4.1, too!

Teacher. Ah, yes. My hope now becomes pious! Anyway, let's begin with the cone equations, and see if from them we can come up with an appropriate box. In Figure 4.1, let's say that the ellipse has equation

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}=1,$$

which homogenizes to the cone

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}=z^2.$$

Then one hyperbola is

$$\frac{x^2}{a^2}-\frac{y^2}{b^2}=1,$$

and this homogenizes to the cone

$$\frac{x^2}{a^2}-\frac{y^2}{b^2}=z^2.$$

The other hyperbola is

$$-\frac{x^2}{a^2}+\frac{y^2}{b^2}=1,$$

and this homogenizes to

$$-\frac{x^2}{a^2}+\frac{y^2}{b^2}=z^2.$$

Suppose the ellipse fits perfectly in standard position on the top face of some rectangular coordinate box. Let's take the edges of this face to have lengths $2a$ in the x-direction, and $2b$ in the y-direction. If we assume that the right face lies in the plane $x = a$, then algebraically that means dehomogenizing at $x = a$. If we do this to the cone

$$\frac{x^2}{a^2}-\frac{y^2}{b^2}=z^2$$

obtained from homogenizing the "horizontal" hyperbola in Figure 4.1, we get:

$$\frac{y^2}{b^2}+z^2=1.$$

If this ellipse is to fit perfectly on the right face of the box—the face contained in the plane $x = a$—then the box must measure $2b$ in the y-direction (which it does!) and $2 \cdot 1$ in the z-direction. We've now tied our hands: the box must have edge lengths $(2a, 2b, 1)$. We hope that the remaining ellipse fits correctly on the remaining face. That is, it should measure $2a$ in the x-direction and $2 \cdot 1$ in the z-direction. Dehomogenizing the last cone,

$$\frac{-x^2}{a^2} + \frac{y^2}{b^2} = z^2,$$

at $y = b$ gives

$$\frac{x^2}{a^2} + \frac{z^2}{1^2} = 1.$$

This ellipse is indeed $2a$ by $2 \cdot 1$, and therefore does fit perfectly on the face contained in the plane $y = b$. It means we can sketch this picture:

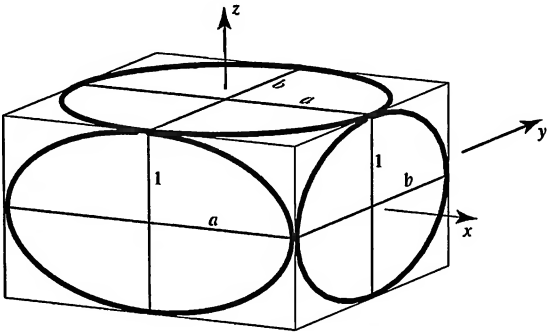


FIGURE 4.2. The central projections of our cones all fit perfectly on the faces of an appropriate rectangular box.

In general, one can dehomogenize using any plane not containing the origin. If the plane is parametrized by $x_i = a_i s + b_i t + c_i$ $(i = 1, \ldots, 3$; not all $c_i = 0)$, then putting x_i into the cone's equation $H(x_1, x_2, x_3) = 0$ gives a quadratic equation in the variables s and t and that, in turn, defines a conic in that plane. This generalizes—H can be a homogeneous polynomial of any degree.

Paradise Regained

Teacher. Let's capitalize on our good fortune. I'm going to sketch the disk version of Figure 4.1 as well as Figure 4.2, adding shading to both. That way we can easily compare

them. Also, let's use a circle instead of an ellipse, purely to make everything visually more symmetric. Here they are, with the horizontal x-axis and vertical y-axis included:

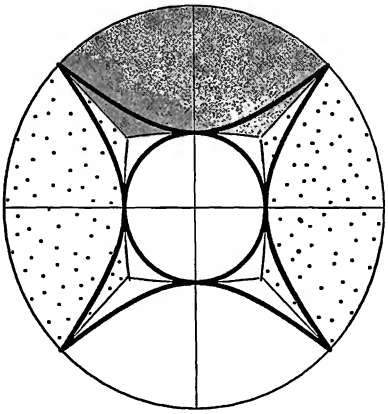


FIGURE 4.3. Mapping a circular version of Figure 4.1 into the disk model and adding shading produces a picture like this. The shading corresponds to that in Figure 4.4.

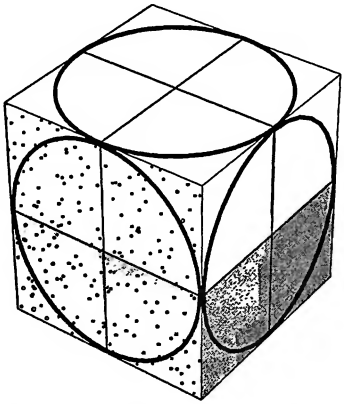


FIGURE 4.4. The flat faces and handy vocabulary of a cube ("vertices," "edges," "faces" and so on) make it a good surface on which to see and talk about the three canonical views of conics—all on an equal footing. On the cube, the constant curvature of a sphere is redistributed to zero on faces, and "infinite curvature" on edges, where faces make a right turn.

For those who want to see things on a sphere...

Consider the cube in Figure 4.4 to be defined by $|x| = 1, |y| = 1, |z| = 1$. Now add in the sphere $x^2 + y^2 + z^2 = 2$. This sphere intersects the cube in six circles. The cube's three "square girths" through the circle centers centrally project to the sphere's three great coordinate circles defined by $x = 0$, $y = 0$ and $z = 0$. The three diagonal girths of the cube—tangent to the heavy circles and only partially drawn—centrally project to the great circles defined by $x = y$, $x = z$ and $y = z$.

Philosopher. Except for the shading, this is beautifully symmetric!

Teacher. After seeing how things correspond in Figure 4.3 and Figure 4.4, one could of course elect to remove the shading.

Philosopher. Are you saying that Figure 4.4 is a version of my original Figure 3.1 [p. 45]?

Teacher. Yes—Figure 4.4 is Figure 3.1 with points at infinity added, and then put in an especially symmetric form.

Student. I like that cube picture—it's clear and easy to understand.

Teacher. Let's look at some correspondences. The slightly-bulging square in Figure 4.3 corresponds to the rectangle of Figure 3.1 or Figure 4.1. The square's interior is white and corresponds to the top face of Figure 4.4, which is also white.

In Figure 4.1 one can look at a "quadrilateral branch"—the part of the quadrilateral surrounding a particular hyperbola branch. The left and right quadrilateral branches (plus the two missing points at infinity that ought to be there to connect the branches) share two opposite sides with the two "up-down" quadrilateral branches. We see these quadrilaterals, with their points at infinity, in Figure 4.3. The left half of the horizontal quadrilateral's interior is speckled, and the right half is speckled and shaded. One can think of going from Figure 4.3 to Figure 4.4 by stretching the disk model's central white square over the top face of the cube and, with appropriate massaging, pasting the white speckled region of the disk to the white speckled region of the cube's left face, pasting the shaded speckled part in the disk to the back side of the cube. We don't see that directly, but since opposite cube-face points are identified, we do in fact see it as the bottom half of the cube's southwest face.

Philosopher. I think I see what you're up to! Continuing, there's that third quadrilateral in Figure 4.1 whose interior appears as the two nonspeckled shaded regions in the disk picture, and which similarly gets pasted over the top halves of the southeast face and its hidden opposite.

Student. But we see it because the hidden points' opposites are visible as the bottom half of the cube's southeast face.

Teacher. Yes. Notice that of the cube's six faces, three are visible and three are hidden. We've so far mentioned two of the hidden faces. The last hidden face is the bottom of the cube—the antipode of the top face.

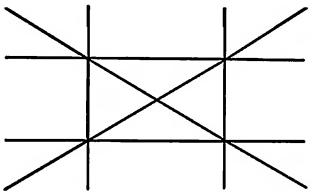
Chapter Summary

This chapter pushes visual unity between ellipse and hyperbolas to a logical extreme. The central idea of the chapter is summed up in Figure 4.4, in which all the cube's faces are equivalent. The cube is a handy version of a sphere, not only because of established terms such as "vertex," "edge" and "face," but also because its flat faces and straight edges make visualizing easy. Essentially, a sphere's visual-cortex-gobbling curvature is redistributed to a few symmetrically-placed right-angle turns.

Exercises

1. Is it possible for two antipodal ellipses appearing in Figure 4.2 on p. 60 to have distinct foci in the x-direction, two other antipodal ellipses there to have distinct foci in the y-direction, and the remaining two to have distinct foci in the z-direction? Why or why not?

2. The sides of the three bounding quadrilaterals in Figure 4.1 can be extended to lines, giving this picture:



Sketch an analogue of Figure 4.4 (p. 61), showing only these six lines. Make the rectangle defined by the above horizontal and vertical lines correspond to the top and bottom cube faces.

3. **The $6 \cdot 4 = 8 \cdot 3 = 24 \cdot 1 = 48 \cdot \frac{1}{2}$ phenomenon:** If we ignore shading, then all six inked faces of the cube in Figure 4.4 are identical, and this means you can "tile" the cube with copies of any one face. But this is only one of several natural tilings of the cube. For example, notice that the cube is composed of eight smaller cubes, each one having three exposed faces. All eight of these "3-face tiles" are mutually congruent and can thus also tile the cube—we can write 24 as $8 \cdot 3$. But even further, all 24 of these exposed 1×1 faces are congruent to each other ($24 = 24 \cdot 1$), and it's not hard to see an even finer, $48 \cdot \frac{1}{2}$-tiling using triangles. *A sphere analogue*: A sphere centered at (0,0,0) is divided into eight congruent spherical triangles by the three coordinate planes, and each of these triangles defines a spherical tile. Sketch a sphere version of Figure 4.4 on a Ping-Pong ball, and observe the spherical $6 \cdot 4 = 8 \cdot 3 = 24 \cdot 1 = 48 \cdot \frac{1}{2}$ phenomenon on your creation.

4. Using the idea given in the shaded box on p. 60, homogenize the curve $xy = 1$; then dehomogenize at a non-coordinate plane so a point originally at infinity becomes the new origin.

5. Find a parametrization of the plane P: $x + y + z = 1$ in terms of s and t, as suggested in the box on p. 60. Then find, in terms of s and t, the equation of the conic in which P intersects the cone $x^2 + y^2 = z^2$. As a check, use Maple to sketch the locus of your equation, comparing that with what your "mind's eye" tells you.

5

The Other Foci: A Well-Kept Secret

in which the trio discovers two new focal points

Complex Conics

Philosopher. I'm probably pushing my luck, but I'd sure be interested in conquering another problem that has bothered me for a long time.

Teacher. What's that?

Philosopher. Why is it that ellipse foci act so differently from hyperbola foci?

Teacher. What do you mean?

Philosopher. Take the hyperbola

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1,$$

for example. No matter how I change its shape by altering a and b, its foci stay on the x-axis. Ellipse foci don't act that way—they lie on whichever diameter happens to be longer. If $a > b$, the foci of the ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

are on the x-axis. But if you change the ellipse by decreasing a, the foci come together when $a = b$, and then live on the y-axis when $a < b$. Focus locations depend upon the shape of the ellipse.

Teacher. I'm not sure about that. It seems to me that ellipses and hyperbolas are somewhat similar in that regard. For an ellipse, the foci lie on the longer axis of the basic rectangle surrounding it. A hyperbola? Well, we've seen that the usual box whose diagonal lines are the asymptotes is *not* the hyperbola's fundamental surrounding box—rather, it's the quadrilateral that we've just discovered. Now, if the hyperbola opens to the left and right, then the vertical quadrilateral sides are finite, while the nonvertical sides are infinite and therefore always longer, no matter how we change a and b.

Philosopher. That's pretty ingenious, and maybe you're right. But I still have a problem with the formulas for the focal length of an ellipse versus that of a hyperbola. The focal

length of the ellipse

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}=1$$

is $\sqrt{a^2-b^2}$—but that's *only* if the a-axis happens to be longer than the b-axis. Otherwise, the formula reads $\sqrt{b^2-a^2}$. Give me a break! That sounds exactly like mathematics instead of Mathematics.

Teacher. I have to agree. It reminds me of how, before complex numbers were introduced, $x^2-1=0$ had solutions, but the assumption was that $x^2+1=0$ did not.

Philosopher. Hyperbola focal lengths behave differently, because for them the focal length is *always* $\sqrt{a^2+b^2}$ This applies to both

$$\frac{x^2}{a^2}-\frac{y^2}{b^2}=1 \quad \text{and} \quad -\frac{x^2}{a^2}+\frac{y^2}{b^2}=1,$$

and is completely independent of which of a or b is larger.

Teacher. Well, you certainly make a case that ellipses and hyperbolas don't seem to be acting like blood brothers here.

Student. Have we reached the end of the road on seeking unity between them?

Teacher. We won't go down that easily! Let's begin by looking at your first problem: "The focal length of an ellipse is $\sqrt{a^2-b^2}$, but only if the a-axis is longer than the b-axis." Suppose we simply drop the *proviso* "only if the a-axis is longer than the b-axis," and *insist that the formula always hold.* What then?

Student. We get nonsense, because when $a<b$, then $\sqrt{a^2-b^2}$ is imaginary!

Teacher. Imaginary "distances" aren't something one considers every day, but physicists routinely use the idea of "separation squared" as a generalization of distance squared, and separation squared can be negative, meaning that separation is imaginary. In fact, all this is part of our real world.

Student. What do you mean? What's a real-world example of a negative separation squared?

Teacher. I'm talking about "distances" between points in space-time! The metric in space-time regards time as imaginary. Here's a specific example:

In Figure 5.1, the heavily-drawn line segment is a *space-time diagram* of a point moving to the right in 1-dimensional space. Think of the plane together with a space-time diagram as a VCR tape together with the magnetic signal on the tape. In our case, a simple playback machine is a piece of cardboard in which we've cut out a narrow horizontal slit. That slit is a one-dimensional "TV screen." We insert our analogue of tape-plus-magnetic signal (that is, the plane with a space-time diagram drawn on it) by placing it under the slit (holding the slit fixed), and we turn the player on by moving the space-time diagram downward at a constant speed. This is analogous to tape passing over the fixed VCR playback head. One sees in the slit a point moving to the right. The point (0, 0) represents the "here-now" having space coordinate 0 at time 0.

Student. So staring at that point would be analogous to looking at a "freeze frame" image on the TV screen?

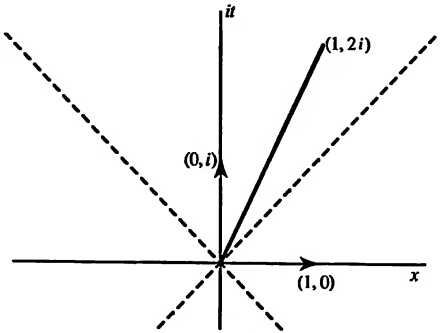


FIGURE 5.1. In this simplified "space-time diagram," space is one-dimensional, the vertical axis representing time. The dotted lines form the "light cone" through (0, 0). Any vertical line can be regarded as the history of a stationary point. The heavily-drawn line segment represents a point popping into existence at time $t = 0$, moving rightward at constant speed, and disappearing when $it = 2i$.

Interactive Applets

For an interactive version of Figure 5.1, see Chapter Applet 3 in the CD-ROM.

Teacher. Exactly. As the space-time diagram is dragged downward, the point in the slit-shaped "TV screen" moves at a constant speed to the right, disappearing at the here-now $(1, 2i)$. To record that the point continues to exist but remains at rest, we'd need to draw in a rising vertical ray starting from $(1, 2i)$. The separation squared s^2 of the heavy line segment in Figure 5.1 is calculated using the ordinary point-to-point distance formula:

$$s^2 = 1^2 + (2i)^2 = -3.$$

The separation s is therefore $\sqrt{3}i$.

Convention. Throughout this book, the term "imaginary" will always mean *pure* imaginary.

Definition 5.1. Any plane having one real coordinate axis and one imaginary coordinate axis, with "distance-squared" s^2 between points (a, ib) and (c, id) given by $s^2 = (a - c)^2 - (b - d)^2$, is called a **Minkowski plane.**

Philosopher. But where do imaginary quantities fit in with our conics? All our conics live in the real world!

Teacher. We need to expand our idea of space. We did this once when we extended the Euclidean plane by adding points at infinity to get the projective plane. Now we're going

to extend Euclidean space in a different way—extending real lines to "complex lines." The more customary term for $\mathbb{C}$ is, of course, "complex *plane*," but actually that term has some drawbacks. We may think of extending the field $\mathbb{R}$ (which functions as the line of scalars of the vector space $\mathbb{R}^2$) to the field $\mathbb{C}$, serving as the line of scalars of the vector space $\mathbb{C}^2$; the $\mathbb{C}$-vector space $\mathbb{C}^2$ has complex dimension 2.

Hermann Minkowski
1864–1909

The Minkowski plane is well named: when "space-time" is mentioned, most people think of Einstein—but the idea is actually due to Minkowski. Einstein was initially skeptical of what seemed like just a formal abstraction. To him, it seemed to accomplish little. He gradually came to appreciate its fundamental significance, and eventually used it in a central way in his subsequent general theory of relativity.

Philosopher. In terms of real dimensions, $\mathbb{C}^2$ is like $\mathbb{R}^4$

Teacher. Yes. In $\mathbb{C}^2$, coordinates (x, y) are ordered pairs of complex numbers, and each complex number can be looked at as an ordered pair of reals, its real and imaginary parts.

Now for the big question: "*In $\mathbb{C}^2$, what is a conic?*" The answer happens to be straightforward: if a conic in $\mathbb{R}^2$ is given by some equation of degree 2, say

$$Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0,$$

then this equation also defines a conic in $\mathbb{C}^2$ consisting of all the points in $\mathbb{C}^2$ satisfying this equation. However, the coefficients $A, \ldots, F$ can be complex. For our purposes, we'll usually take these coefficients to be real or imaginary, which, it turns out, means we can get informative peeks into $\mathbb{C}^2$ by looking in simple coordinate 3-D slices.

The locus of a polynomial is very context-dependent. For example in the x-axis, $x = 0$ defines a point, while in the (x, y)-plane, it defines a line, and more generally in $\mathbb{R}^n$, a hyperplane. In $\mathbb{R}^2$, $x^2 + y^2 = -1$ defines the empty set, but in $\mathbb{C}^2$ it defines a real 2-manifold, which turns out to be topologically the same as the complex circle $x^2 + y^2 = 1$.

Looking in 3-D Slices of Four-Space

Philosopher. Visualizing in four dimensions sounds daunting.

Teacher. I agree with you, but by looking in certain three-dimensional slices of $\mathbb{R}^4$, the four-dimensional picture gives up some of its secrets. Let me show you what I mean.

Write

$$x = x_1 + ix_2 \quad \text{and} \quad y = y_1 + iy_2.$$

Then the point (x, y) in $\mathbb{C}^2$ can be regarded as a point with coordinates (x_1, ix_2, y_1, iy_2), and by setting any one of these four coordinates equal to a constant, we get a special 3-space in $\mathbb{R}^4$. Crucial parts of the full conic reveal themselves in certain of these three-dimensional slices.

Student. What is a specific example?

Teacher. Well, for one, you'd be amazed at what a parabola *really* looks like! Nearly everyone has seen only a tiny bit of a parabola. It's a lot like Hawaii, in that we see just a little bit of the submerged mountain range. "Removing the water" hiding the vastness of Hawaii's entire mountain range would be quite an educational experience, and that's true of the parabola, too. So let's "pull the oceanic plug" on the parabola $y = x^2$ and see what we get.

To begin, note that every point of the complex x-axis is mapped via $y = x^2$ to a point in $\mathbb{C}^2$. The complex parabola has two real dimensions in $\mathbb{R}^4$, not one. We can find specific real equations for this set by substituting $x = x_1 + ix_2$ and $y = y_1 + iy_2$ and equating real and imaginary parts:

$$\begin{aligned} y_1 + iy_2 &= (x_1 + ix_2)^2 \\ y_1 + iy_2 &= x_1^2 + 2ix_1x_2 - x_2^2 \\ y_1 &= x_1^2 - x_2^2, \qquad y_2 = 2x_1x_2. \end{aligned} \tag{5.1}$$

Each of the last two equations in (5.1) imposes one dimension-decreasing constraint on $\mathbb{R}^4$, meaning that their simultaneous solution is a real two-dimensional set in $\mathbb{R}^4$ The complex parabola is a surface—a smooth, 2-D graph in $\mathbb{C}^2 = \mathbb{R}^4$. By setting any one of the variables equal to a constant, we get a three-dimensional slice of $\mathbb{R}^4$, and the part of the complex parabola within this 3-slice is usually a real space curve.

For our first 3-D peek, let's set y_2 equal to zero. Then the second equation in (5.1) becomes $x_1x_2 = 0$, which of course implies

$$\text{either } x_1 = 0 \quad \text{or} \quad x_2 = 0.$$

Now if $x_1 = 0$, then the first equation in (5.1) becomes $y_1 = -x_2^2$. That means we're actually assuming both $y_2 = 0$ and $x_1 = 0$, and that of course gives us the (x_2, y_1)-plane in $\mathbb{R}^4$—or, in $\mathbb{C}^2$, the (ix_2, y_1)-plane. Thus we get an ordinary parabola $y_1 = -x_2^2$ in that plane. Similarly, if $x_2 = 0$, then the first equation reduces to $y_1 = x_1^2$. We're now assuming both $y_2 = 0$ and $x_2 = 0$, and this gives us the (x_1, y_1)-plane in $\mathbb{R}^4$. Therefore we see the ordinary parabola $y_1 = x_1^2$ in, of course, the usual real plane.

We can sketch both these curves like this:

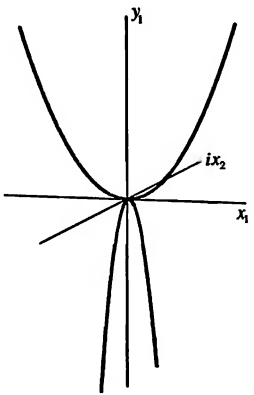


FIGURE 5.2. This is the part of the complex parabola $y = x^2$ in a slice of $\mathbb{C}^2$. In (x_1, ix_2, y_1)-space, we see two parabolas—the usual one in the (x_1, y_1)-plane, and one in the (ix_2, y_1)-plane, opening downward.

Extending the real setting to the complex is like turning on a powerful floodlight, illuminating things previously hidden in the dark. This often answers questions—even sometimes ones we didn't even think of asking. For example, looking in (x_1, ix_2, y_1)-space tells an interesting story about the parabola.

To illustrate what I mean, consider the real parabola $y = x^2$ in the ordinary plane $\mathbb{R}^2$. Now let's translate its graph up and down while holding the axes fixed. When we push it down, we clearly see two "roots,"—points where it crosses the x_1-axis. From this position, slowly elevate the graph; the two roots meet, becoming two points occupying the same position. We can say that this is a point of "multiplicity two." But as we continue elevating the graph, the (x_1, y_1)-screen goes blank, in that we see no roots at all. The morphing process, which is usually so powerful in tracking things, suddenly runs dry and yields no information. Although when the parabola is pushed up to become $y = x^2 + 1$, we can set y equal to zero and get roots $\pm i$, this is in a sense algebraic magic. We're still missing the continuation of the visual morphing process. The screen went blank because the action went off the screen. A 3-D slice serves as a larger screen, and in it we can continue to follow the action.

Now as we've just seen, the part of the complex parabola in our 3-D screen is a space curve consisting of two ordinary-looking parabolas. Let's play a little movie that starts by showing the entire double parabola at a pushed-down state. In the starting scene of

the movie, we see the usual top part of the curve intersecting the x_1-axis in two points, and as we play the movie by slowly elevating the whole curve, we see the roots coalesce. With continued elevating, the bottom parabola, which is oriented perpendicular to the top one, now jumps into action, intersecting the ix_2-axis. We see two pure imaginary roots symmetrically moving apart.

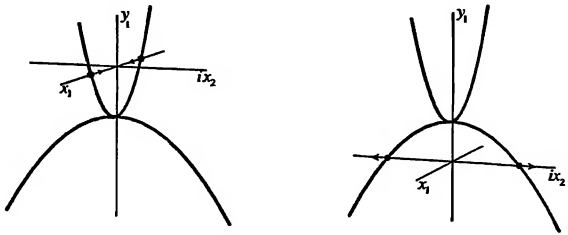


FIGURE 5.3. As the double parabola moves upward, its two intersection points with the complex x-axis symmetrically move toward the origin along the real x_1-axis, then symmetrically move apart along the imaginary ix_2-axis.

A Hoax

Philosopher. This is inspired! Am I right to suspect that this floodlight is going to help me out of the problem about ellipse focal lengths? The standard answer is that the length is $\sqrt{a^2 - b^2}$, provided the a-axis is longer than the b-axis, and you suggested dropping the condition that the a-axis has to be longer than the b-axis.

Teacher. Right. Starting with an ellipse with $a > b$, let's morph it by steadily decreasing a. The horizontally-elongated ellipse therefore becomes less elongated and more nearly circular. Its two foci approach each other, much as they did a minute ago when we pushed up the double parabola and the two roots approached each other. This is algebraically obvious because the foci are at $(\pm\sqrt{a^2 - b^2}, 0)$. When $a = b$, they become $(\pm 0, 0)$, meaning that the center of a circle can be regarded as two coincident foci.

What happens next, however, seems different from the case of the parabola. As we continue to decrease a, the foci now appear to separate along the *real* y-axis. In the parabola example, the roots stayed in the complex x-axis, turning at right angles *within the complex x-axis*. The big question is now: if we turn on the floodlights and look in the larger, complex screen, will some deeper truth reveal itself? We normally see the foci suddenly turning from lying on the x-axis, to lying on the y-axis, but could this actually be a red herring? That is, if we restrict ourselves by looking exclusively in the real screen, then just after the foci on the x-axis coalesce, perhaps the real x-screen goes blank as far as *those particular foci are concerned*, and we might in fact be seeing two *other* foci simultaneously appearing in the real y-axis!

Philosopher. Are you proposing that what everyone sees, as the shape of an ellipse varies from $a > b$ to $a < b$, is really a *ruse*?!

Student. (Excitediy) We could be uncovering a major hoax, perpetrated on the mathematical public by the ease of seeing in the real plane and the difficulty of seeing in four dimensions!

> Start with an ellipse initially looking like a horizontal cigar. As you compress in the x-direction so the ellipse eventually becomes taller than it is wide, the foci on the x-axis appear to run together and then separate along the y-axis. That's a deception. But because it's hard to visualize in four dimensions, we fall for it.

Teacher. We need to do an ellipse version of our parabola experiment. Continuing to write

$$x = x_1 + ix_2 \quad \text{and} \quad y = y_1 + iy_2,$$

we substitute these into the equation of an ellipse, say

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1,$$

expand, and equate real and imaginary parts. We just mimic what we did for the parabola:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

$$\frac{(x_1 + ix_2)^2}{a^2} + \frac{(y_1 + iy_2)^2}{b^2} = 1$$

$$\frac{x_1^2 - x_2^2}{a^2} + \frac{y_1^2 - y_2^2}{b^2} + 2i\left(\frac{x_1x_2}{a^2} + \frac{y_1y_2}{b^2}\right) = 1$$

$$\frac{x_1^2 - x_2^2}{a^2} + \frac{y_1^2 - y_2^2}{b^2} = 1, \qquad \frac{x_1x_2}{a^2} + \frac{y_1y_2}{b^2} = 0. \tag{5.2}$$

As in the parabola example, we set y_2 equal to zero. The second equation of (5.2) then implies $x_1 = 0$ or $x_2 = 0$. As with the parabola example, we get one curve in the (x_1, y_1)-plane (obtained by setting $x_2 = 0$) and another one at right angles to it in the (ix_2, y_1)-plane (obtained by setting $x_1 = 0$). These two curves are

$$\text{the usual real ellipse } \frac{x_1^2}{a^2} + \frac{y_1^2}{b^2} = 1 \text{ in the } (x_1, y_1)\text{-plane;}$$

$$\text{the hyperbola } -\frac{x_2^2}{a^2} + \frac{y_1^2}{b^2} = 1 \text{ in the } (ix_2, y_1)\text{-plane.}$$

Here's the picture:

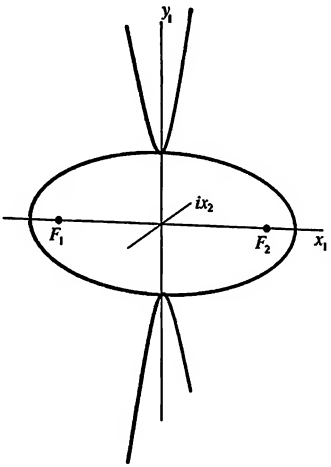


FIGURE 5.4. This three-dimensional slice of a complex ellipse reveals a hyperbola in the Minkowski (ix_2, y_1)-plane. Though we may call it a hyperbola, it's of course still part of the complex ellipse.

Figures 5.5–5.7 below show three shots of an ellipse with its horizontal axis decreasing. As the ellipse becomes more and more circular, the foci approach one another along the x_1-axis, meeting when $a = b$. After that, those foci separate along the ix_2-axis.

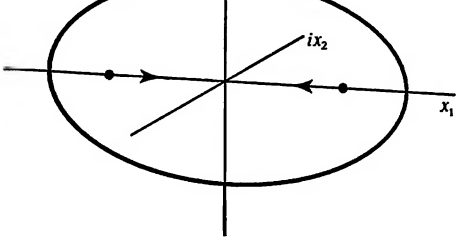


FIGURE 5.5. As the larger semi-axis length a decreases toward b, the foci approach the center of the ellipse.

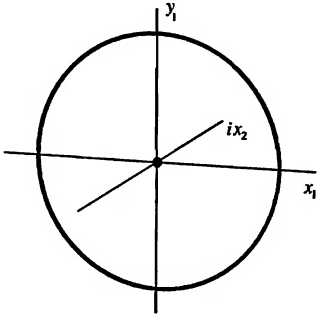


FIGURE 5.6. Here, $a = b$ and the foci have coalesced to a double point. The center of any circle can be thought of as the two coincident foci of an ellipse with equal radii. Ellipse radii are the maximum and minimum distances from the center—just the semi-major and semi-minor axis lengths.

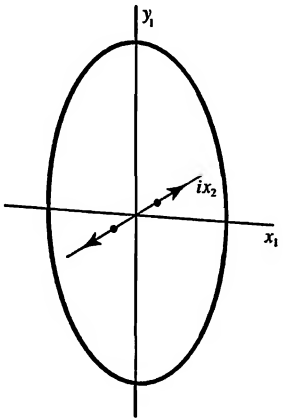


FIGURE 5.7. Here $a < b$, and the x-foci are now imaginary. As a decreases, these foci separate along the ix_2-axis.

Student. Those foci on an imaginary axis sure make things look strange! Do they actually act like foci? Do the distances from the foci to any point on the ellipse add up to $2a$?

Philosopher. ...and what about the reflection property?

Teacher. We obviously need to look at a couple of examples to see if my proposal makes sense. Or—who knows—maybe nonsense!

> "Schrödinger and I both had a very strong appreciation of mathematical beauty, and this appreciation of mathematical beauty dominated all our work. It was a sort of act of faith with us that any equations which describe fundamental laws of Nature must have great mathematical beauty in them. It was like a religion with us. It was a very profitable religion to hold, and can be considered the basis of much of our success."
>
> P.A.M. Dirac, 1977

Consider the ellipse

$$\frac{x^2}{1^2}+\frac{y^2}{2^2}=1$$

which, like the one in Figure 5.7, is taller than it is wide, and let's see if the sum-of-lengths property holds. The separations from the x-foci to a point P on this ellipse should sum to $2a=2$.

> Dirac once said, "The present quantum electrodynamics does not conform to the high standard of mathematical beauty that one would expect for a fundamental physcial theory, and that leads one to suspect that a drastic alteration of basic ideas is still needed."

A focal length of $\sqrt{a^2-b^2}$ puts the proposed x-axis foci at $F_1=(-i\sqrt{3},0)$ and $F_2=(+i\sqrt{3},0)$. Using $P=(1,0)$ as one point on the ellipse, we get for F_1P+PF_2

$$\sqrt{(1-\sqrt{3}i)^2+0^2}+\sqrt{(1+\sqrt{3}i)^2+0^2}=(1-\sqrt{3}i)+(1+\sqrt{3}i)=2.$$

Success!

Philosopher. I really want to agree with you, but there seems to be a sticky point here. You've apparently chosen

$$\sqrt{(1-\sqrt{3}i)^2+0^2}\quad\text{to be}\quad(1-\sqrt{3}i),$$

but what if you choose $-(1-\sqrt{3}i)$, instead? Same question for the other radical. That way, it seems to me, you could get -2, or even $\pm 2i\sqrt{3}$ as F_1P+PF_2.

Teacher. But you could apply the same objections to ordinary real lengths.

Philosopher. True, and we solve the problem by agreeing to take the *positive* square root as the length. But I don't know what one would mean by the "positive square root" if the answer is complex.

Teacher. Your objection is certainly valid, but it does have a nice answer—there *is* a standard definition of "positive" for complex numbers. It's the convention used by popular complex computing engines such as Mathematica, Maple or a calculator with complex number capability.

This picture gives the idea:

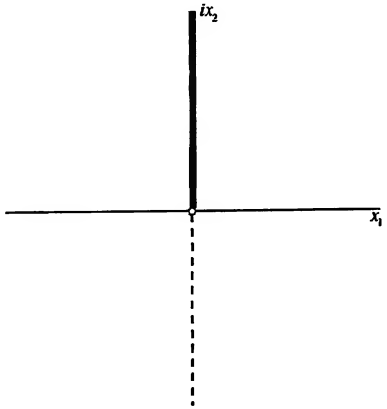


FIGURE 5.8. The set of "positive" complex numbers consists of [shaded area] ∪ [positive vertical ix_2-axis].

Therefore the "positive" square root of any complex number is obtained this way: *if the real part is nonzero, choose the square root having positive real part. If the real part is zero, then choose the square root having positive imaginary part.*

A *lexicographic* or "dictionary" ordering of the real plane is like the ordering of names in a telephone book. It treats an axis—say, the x-axis—as "Last name," and the y-axis as "First name." This mimics alphabetical ordering, in which Smith, James < Smith, John < Smythe, Adam. Notice that this is a "total order"—that is, for any two distinct elements, one is the larger and the other, the smaller.

Here is its full definition:

Let (a, b), (a', b') be two points in $\mathbb{R}^2$. Then $(a, b) < (a', b')$ if and only if either: $a < a'$; or $a = a'$ and $b < b'$.

In Figure 5.8, [the shaded area] ∪ [the positive vertical axis] makes up the set of positive elements relative to this ordering. The positive points are closed under addition but not multiplication—the complex numbers cannot be totally ordered as a field.

Philosopher. That pretty well answers my objection! So that's definitely good news, but I think that we should try a more "arbitrary" point. That is, the symmetric position of (1,0) on the ellipse may be leading to an overly-promising outcome.

Teacher. OK, try the asymmetrically-positioned point $P = (-\frac{1}{2}, -\sqrt{3})$. Then $F_1P + PF_2$ turns out to be

$$\left(1 - \frac{\sqrt{3}}{2}i\right) + \left(1 + \frac{\sqrt{3}}{2}i\right) = 2.$$

Philosopher. You're starting to make a believer out of me! But I'd like to see a general proof.

Teacher. If one can demonstrate a purely algebraic proof of the ordinary sum-of-lengths property, then that very same proof should work even if imaginary quantities are present. This is because the same Pythagorean, point-to-point formula is used for calculating separations as for ordinary distances. Here is such an algebraic proof:

Proof of sum-of-lengths property. Let (x, y) be any point on the real ellipse in $\mathbb{R}^2$ defined by

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1, \tag{5.3}$$

where $a > b$. Let $L_i(x)$ be the distance from F_i to (x, y). Then, for example,

$$(L_1(x))^2 = \left(x - \left(-\sqrt{a^2 - b^2}\right)\right)^2 + y^2 \tag{5.4}$$

Solving (5.3) for y^2, substituting into (5.4) and rewriting slightly gives

$$(L_1(x))^2 = \left(a + \frac{\sqrt{a^2 - b^2}}{a} \cdot x\right)^2$$

from which we get

$$L_1(x) = a + \frac{\sqrt{a^2 - b^2}}{a} \cdot x. \tag{5.5}$$

Similarly,

$$L_2(x) = a - \frac{\sqrt{a^2 - b^2}}{a} \cdot x. \tag{5.6}$$

From this we see at once that $L_1(x) + L_2(x) = 2a$.

The Reflection Property

Philosopher. I initially greeted the idea of a second set of foci as nonsense, but it's now starting to seem like a "natural." To be thoroughly convincing, however, the new set of foci

must have the reflection property, too. That is, the line segments $\overline{F_1P}$ and $\overline{F_2P}$ must both make the same angle with the line normal to the ellipse at P.

Teacher. Of course one must decide how to define and compute angles in $\mathbb{C}^2$. Let's continue with the same algebraic approach that we used in defining distances. One natural way is to begin with a familiar and powerful tool for computing angles in a real vector space—the dot product:

$$\mathbf{v} \cdot \mathbf{w} = \|v\| \cdot \|w\| \cdot \cos\alpha. \tag{5.7}$$

Here α is either angle between vectors $\mathbf{v}$ and $\mathbf{w}$. If $\mathbf{v} = (v_1, v_2)$ and $\mathbf{w} = (w_1, w_2)$, then

$$\|\mathbf{v}\| = \sqrt{v_1^2 + v_2^2} \quad \text{and} \quad \|\mathbf{w}\| = \sqrt{w_1^2 + w_2^2}.$$

But $\mathbf{v} \cdot \mathbf{w}$ has another form, too:

$$\mathbf{v} \cdot \mathbf{w} = v_1 \cdot w_1 + v_2 \cdot w_2. \tag{5.8}$$

Together, (5.7) and (5.8) are very useful, because they say that

$$\|\mathbf{v}\| \cdot \|\mathbf{w}\| \cdot \cos\alpha = v_1 \cdot w_1 + v_2 \cdot w_2,$$

which leads to

$$\cos\alpha = \frac{v_1 \cdot w_1 + v_2 \cdot w_2}{\sqrt{v_1^2 + v_2^2}\sqrt{w_1^2 + w_2^2}}. \tag{5.9}$$

This is a purely algebraic expression for the angle between vectors $\mathbf{v}$ and $\mathbf{w}$, and *we can use it to obtain* α *in* $\mathbb{C}^2$. Of course (5.9) immediately generalizes to $\mathbb{R}^n$ and $\mathbb{C}^n$, though we don't need it.

Philosopher. I do see a little problem here, because $\cos\alpha$ doesn't completely determine α. Within any half-open interval of length 2π there are generally two α's for any $\cos\alpha$.

Teacher. You are of course right, and we need to deal with that. Let's first see what is going on physically. Think of a photon reflecting from a surface as like a tennis ball bouncing off a tennis court's surface. The ideal trajectory of the tennis ball is completely determined, and is, of course, real. If we express that unique real solution algebraically, we can directly extend the solution to the complex setting.

When a tennis ball bounces off a smooth surface at a point P, we can think of the surface as breaking up the incident velocity vector into components tangential and normal to the surface. The reflected velocity vector has the same tangential component as the incident vector, but its normal component is opposite to that of the incident vector. This implies that the angle of incidence equals the angle of reflection. In Figure 5.9, we've turned the incident vector around so that not only are the incident and reflected angles equal, but all vectors emanate from the point P.

Now the incidence and reflection angles can be measured from either the normal vector $\mathbf{n}$ or from the tangent vector $\mathbf{t}$. Measured from $\mathbf{n}$, "angle of incidence equals the angle of reflection" translates to $\alpha = -\beta$ in the figure. Measured from $\mathbf{t}$, it translates to $\phi = \pi - \psi$.

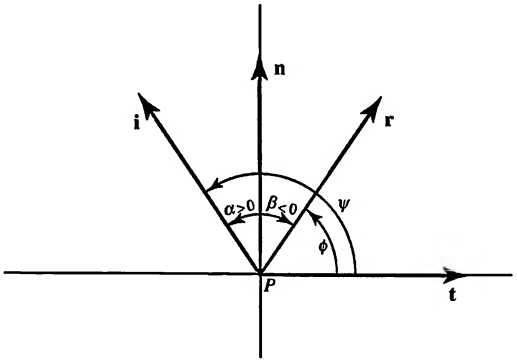


FIGURE 5.9. In terms of angles, "the angle of incidence equals the angle of reflection" means $\alpha = -\beta$. This further translates to the two conditions $\cos\alpha = \cos\beta$ and $\cos\phi = -\cos\psi$. These four cosines are easily computed algebraically, using dot products; the formulas—which are algebraic—allow one to translate to the complex setting, where we can establish the general incidence-reflection property.

Our basic computational tool is the dot product, and it's easy to compute the slightly less-informative cosines of the angles in these two equalities. Taking appropriate dot products leads us to two facts: first, $\cos\alpha = \cos(-\beta) = \cos\beta$; second, $\cos\phi = \cos(\pi - \psi) = -\cos(\psi)$. Is this enough information to conclude that $\alpha = -\beta$? The good news is, *yes*. Here's why: first, $\cos\alpha = \cos\beta$ produces two candidates

$$\alpha = \pm\beta (\mathrm{mod}\, 2\pi), \tag{5.10}$$

and $\cos\phi = -\cos(\psi)$ also produces two candidates, $\phi = \pi \pm \psi (\mathrm{mod}\, 2\pi)$. Now $\phi = \frac{\pi}{2} + \beta$ and $\psi = \frac{\pi}{2} + \alpha$, and a simple argument shows that $\phi = \pi \pm \psi (\mathrm{mod}\, 2\pi)$ can also be expressed as

$$\alpha = -\beta (\mathrm{mod}\, 2\pi) \quad \text{or} \quad \alpha = (\beta + \pi) (\mathrm{mod}\, 2\pi). \tag{5.11}$$

The only common choice in (5.10) and (5.11) is $\alpha = -\beta (\mathrm{mod}\, 2\pi)$, which means that $\alpha = -\beta$ in Figure 5.9.

Philosopher. So in the real case, to verify the reflection property, it's enough to verify that $\cos\alpha = \cos\beta$, and $\cos\phi = -\cos(\psi)$. That's fine, but how do you get these cosines?

Teacher. Just read off these dot products from Figure 5.9:

$$\cos\alpha = \frac{\mathbf{i}\cdot\mathbf{n}}{\|\mathbf{i}\|\cdot\|\mathbf{n}\|}, \qquad \cos\beta = \frac{\mathbf{r}\cdot\mathbf{n}}{\|\mathbf{r}\|\cdot\|\mathbf{n}\|};$$

$$\cos\phi = \frac{\mathbf{i}\cdot\mathbf{t}}{\|\mathbf{i}\|\cdot\|\mathbf{t}\|}, \qquad \cos\psi = \frac{\mathbf{r}\cdot\mathbf{t}}{\|\mathbf{r}\|\cdot\|\mathbf{t}\|}.$$

Philosopher. Those four quotients are certainly algebraic! Because of that, I assume we extend to the complex setting by maintaining their exact algebraic form...

Teacher. Yes, that's the method that has worked so well for us!

Philosopher. Well, I want to see how everything pans out with a specific example.

Teacher. It's that time. Once again consider our ellipse

$$\frac{x^2}{1^2}+\frac{y^2}{2^2}=1$$

with imaginary x-foci at $(\mp i\sqrt{3},0)$, and let's continue to use the not-too-symmetric point $P=(-\frac{1}{2},-\sqrt{3})$ on this ellipse. Then the vector from P to F_1 is $\mathbf{i}=(\frac{1}{2}-\sqrt{3}i,\sqrt{3})$ and the vector from P to F_2 is $\mathbf{r}=(\frac{1}{2}+\sqrt{3}i,\sqrt{3})$. We of course need a normal vector to the ellipse at P, and for this we continue to define and compute algebraically. Implicitly differentiating our ellipse equation

$$\frac{x^2}{1^2}+\frac{y^2}{2^2}=1$$

gives $y'=-\frac{4x}{y}$, so a normal vector at (x,y) has slope $\frac{y}{4x}$. At $P=(-\frac{1}{2},-\sqrt{3})$, this slope is $\frac{\sqrt{3}}{2}$, so we can choose our normal vector to be, say, $\mathbf{n}=(2,\sqrt{3})$. Any tangent vector $\mathbf{t}$ must have dot product 0 with $\mathbf{n}$; a natural choice is

$$\mathbf{t}=(\sqrt{3},-2). \tag{5.12}$$

Now that we have four appropriate vectors, we submit them to the acid test. We get:

$$\cos\alpha=\frac{\mathbf{i}\cdot\mathbf{n}}{\|\mathbf{i}\|\cdot\|\mathbf{n}\|}=\frac{4}{\sqrt{7}};$$

$$\cos\beta=\frac{\mathbf{r}\cdot\mathbf{n}}{\|\mathbf{r}\|\cdot\|\mathbf{n}\|}=\frac{4}{\sqrt{7}};$$

$$\cos\phi=\frac{\mathbf{i}\cdot\mathbf{t}}{\|\mathbf{i}\|\cdot\|\mathbf{t}\|}=\frac{-3i}{\sqrt{7}};$$

$$\cos\psi=\frac{\mathbf{r}\cdot\mathbf{t}}{\|\mathbf{r}\|\cdot\|\mathbf{t}\|}=\frac{3i}{\sqrt{7}}.$$

We therefore have $\cos\alpha=\cos\beta$, and $\cos\phi=-\cos\psi$, so we've verified that the angle of incidence equals the angle of reflection in this imaginary-foci example.

Philosopher. Sweet!

Student. Congratulations!

Nearly everyone assumes an ellipse or hyperbola has only two foci. Each actually has *four*.

Student (After thinking a moment). But I see some problems here: on one hand, you have the cosines of α and β being $\frac{4}{\sqrt{7}}$, which according to my calculator is about 1.512, and that's greater than 1. Even worse, you've got the cosines of ϕ and ψ both being imaginary. None of this looks right—the cosine of any angle is real and lies between $+1$ and -1.

Teacher. That's true only when we restrict ourselves to real angles. But one can take the cosine of complex numbers, too, and the results are generally complex. In fact, the cosine function maps $\mathbb{C}$ onto $\mathbb{C}$.

Student. How do you take the cosine of a complex number?

Teacher. Once again, the idea is to define and compute algebraically—or, in this case, analytically. The usual power series for cosine generalizes directly to complex numbers:

$$\cos z = \frac{z^0}{0!} - \frac{z^2}{2!} + \frac{z^4}{4!} - \cdots.$$

This can be shown to converge for any $z \in \mathbb{C}$.

Student. OK, what specific angle has cosine equal to $\frac{4}{\sqrt{7}}$? or $\frac{3i}{\sqrt{7}}$? or $\frac{-3i}{\sqrt{7}}$?

Teacher. It's easy to answer such numerical questions by asking, say, Mathematica or Maple, both of which calculate principal values. For example, we get

$$\operatorname{Arccos}\left(\frac{4}{\sqrt{7}}\right) = \frac{i}{2}\ln 7,$$

which is approximately $0.972955i$. Also,

$$\operatorname{Arccos}\left(\frac{3i}{\sqrt{7}}\right) = \frac{\pi}{2} - \frac{i}{2}\ln 7, \quad \text{and} \quad \operatorname{Arccos}\left(\frac{-3i}{\sqrt{7}}\right) = \frac{\pi}{2} + \frac{i}{2}\ln 7,$$

which are about $1.570796 \mp 0.972955i$, respectively.

Student. A *complex* angle? What does that mean? How could you rotate by $0.972955i$ or by $1.570796 \mp 0.972955i$?

Teacher. We answer that by once again using the power of linear algebra! Let's first consider rotation by a *real* angle. This is given by the rotation matrix

$$R_\theta = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}.$$

R_θ rotates any row vector $\mathbf{v}$ in $\mathbb{R}^2$ by the real angle θ about the origin—that is, $\mathbf{v} \to \mathbf{v}R_\theta$. Now R_θ can actually be used to make a natural "movie" of rotation, in which a vector $\mathbf{v}$ swings around its base point: just follow the image of $\mathbf{v}$ under R_θ as θ varies from 0 (the beginning of the movie) to θ (the end). This extends directly to the complex setting: any vector $\mathbf{v}$ in $\mathbb{C}^2$ gets rotated through an angle z by applying

$$R_z = \begin{pmatrix} \cos z & \sin z \\ -\sin z & \cos z \end{pmatrix}$$

to it. The vector $\mathbf{v}R_z$ varies continuously with z, so one can make a movie of turning a vector through a complex angle. There is the usual difference, however, between vary-

ing in $\mathbb{R}$ and varying in $\mathbb{C}$. In $\mathbb{R}$, paths are lines, being constrained to lie in $\mathbb{R}$. But in $\mathbb{C}$, there are many different paths connecting two points of $\mathbb{C}$, and this extra freedom of course adds richness to complex geometry and analysis. Now one could select a straight-line path between 0 and the movie's ending-point z to get a canonical turning movie, but another natural possibility is replacing that straight line—the hypotenuse of a coordinate right triangle—by the triangle's horizontal and vertical legs. That makes the movie into a two-parter—a turn through a real angle, followed by a turn through a pure imaginary angle.

Philosopher. I know we demonstrated in our example that the incident and reflected angles are the same, but it would be nice to see, for example, that $R_{i/2\ln 7}$ actually does rotate the incident ray into the normal, and also rotates the normal into the reflected ray. If we plug **i** into the rotation matrix, do we in fact get **n**?

Teacher. If **i**, **n**, and **r** all have the same Pythagorean length, then we should. So let's first "normalize" these vectors, and then rotate. We should get agreement. Now in the complex setting we're using the same algebraic formulas as in the real setting, so in normalizing, we divide by the square root of the Pythagorean sum-of-squares.

> The familiar definition of norm in $\mathbb{C}^n$ uses complex conjugates:
>
> $$|(x_1, \ldots, x_n)| = \sqrt{x_1\overline{x}_1 + \cdots + x_n\overline{x}_n},$$
>
> and it's designed to ensure a real outcome. By using the "algebraic length"
>
> $$\|(x_1, \ldots, x_n)\| = \sqrt{x_1^2 + \cdots + x_n^2}$$
>
> instead of this, we drop the requirement that distances be real, and gain the fact that the algebra is exactly the same for $\mathbb{C}^n$ as for $\mathbb{R}^n$.

Here are the norms of **i**, **n**, and **r**:

$$\|\mathbf{i}\| = \sqrt{\left(\frac{1}{2} + \sqrt{3}i\right)^2 + (\sqrt{3})^2}, \quad \text{which simplifies to } 1 + i\frac{\sqrt{3}}{2};$$

$$\|\mathbf{n}\| = \sqrt{2^2 + (\sqrt{3})^2} = \sqrt{7};$$

$$\|\mathbf{r}\| = \sqrt{\left(\frac{1}{2} - \sqrt{3}i\right)^2 + (\sqrt{3})^2} = 1 - i\frac{\sqrt{3}}{2}.$$

Dividing **i**, **n**, and **r** by these respective norms results in:

$$\mathbf{I} = \frac{\mathbf{i}}{\|\mathbf{i}\|} = \frac{1}{7}(8 + 3\sqrt{3}i, 4\sqrt{3} - 6i);$$

$$\mathbf{N} = \frac{\mathbf{n}}{\|\mathbf{n}\|} = \frac{1}{\sqrt{7}}(2, \sqrt{3});$$

$$\mathbf{R} = \frac{\mathbf{r}}{\|\mathbf{r}\|} = \frac{1}{7}(8 - 3\sqrt{3}i, 4\sqrt{3} + 6i).$$

Our rotation matrix is

$$R_{i/2\ln 7} = \begin{pmatrix} \cos\left(\frac{i}{2}\ln 7\right) & \sin\left(\frac{i}{2}\ln 7\right) \\ -\sin\left(\frac{i}{2}\ln 7\right) & \cos\left(\frac{i}{2}\ln 7\right). \end{pmatrix}.$$

We can let Maple simplify the entries: the command **convert(cos(I/2*log(7)),exp);** gives

$$\cos\left(\frac{i}{2}\ln 7\right) = \frac{4}{\sqrt{7}};$$

one similarly finds that

$$\sin\left(\frac{i}{2}\ln 7\right) = \frac{3i}{\sqrt{7}}.$$

Therefore our matrix is

$$R_{i/2\ln 7} = \frac{1}{\sqrt{7}}\begin{pmatrix} 4 & 3i \\ -3i & 4 \end{pmatrix}.$$

Substituting, we see that, indeed,

$$\mathbf{I}R_{i/2\ln 7} = \mathbf{N}; \qquad \mathbf{N}R_{i/2\ln 7} = \mathbf{R}.$$

Philosopher. Success again! All that remains is a general proof of the reflection property.

Teacher. Yes, and once again, an algebraic proof for conics over $\mathbb{R}$ also works over $\mathbb{C}$. Here is such a proof.

Proof of the reflection property. Write

$$\begin{aligned} (x, y) &= (a\cos t, b\sin t); \\ \mathbf{i} &= (a\cos t + \sqrt{a^2 - b^2}, b\sin t); \\ \mathbf{r} &= (a\cos t - \sqrt{a^2 - b^2}, b\sin t); \\ \mathbf{n} &= (b\cos t, a\sin t); \\ \mathbf{t} &= (a\sin t, -b\cos t). \end{aligned}$$

After some algebra, we find

$$\begin{aligned} \|\mathbf{i}\| &= a + \sqrt{a^2 - b^2}\cos t; \\ \|\mathbf{r}\| &= a - \sqrt{a^2 - b^2}\cos t; \end{aligned}$$

$$\|\mathbf{n}\| = \sqrt{a^2 \sin^2 t + b^2 \cos^2 t};$$
$$\|\mathbf{t}\| = \sqrt{a^2 \sin^2 t + b^2 \cos^2 t}.$$

After a bit more algebra, we get

$$\cos\alpha = \frac{\mathbf{i}\cdot\mathbf{n}}{\|\mathbf{i}\|\cdot\|\mathbf{n}\|} = \frac{b}{\sqrt{a^2 \sin^2 t + b^2 \cos^2 t}};$$
$$\cos\beta = \frac{\mathbf{r}\cdot\mathbf{n}}{\|\mathbf{r}\|\cdot\|\mathbf{n}\|} = \frac{b}{\sqrt{a^2 \sin^2 t + b^2 \cos^2 t}};$$
$$\cos\phi = \frac{\mathbf{i}\cdot\mathbf{t}}{\|\mathbf{i}\|\cdot\|\mathbf{t}\|} = \frac{-(\sin t)\sqrt{a^2 - b^2}}{\sqrt{a^2 \sin^2 t + b^2 \cos^2 t}};$$
$$\cos\psi = \frac{\mathbf{r}\cdot\mathbf{t}}{\|\mathbf{r}\|\cdot\|\mathbf{t}\|} = \frac{(\sin t)\sqrt{a^2 - b^2}}{\sqrt{a^2 \sin^2 t + b^2 \cos^2 t}}.$$

Therefore $\cos\alpha = \cos\beta$, and $\cos\phi = -\cos\psi$.

Chapter Summary

Superficially, both the formulas and pictures for the foci of an ellipse look quite different from those of a hyperbola. It is realized that once again, this problem arises from using an overly-restricted viewing screen. In broadening our screen, it turns out that $\mathbb{P}^2(\mathbb{R})$ isn't the only useful extension of $\mathbb{R}^2$. Moving from the real to the complex context paves the way to solving a whole array of problems that Philosopher's symmetry- and aesthetics-driven radar picks up. In this chapter, we see how it leads to unifying the center-to-focus distance formula for ellipse versus hyperbola, and this uncovers a hitherto unsuspected, new pair of foci. It is discovered that a pair of foci exists on each principal axis—not merely on the longer one. We establish that these new foci act exactly as they should. For example, the sum-of-lengths property holds and the incident and reflected angles are equal.

Exercises

1. Fill in the argument leading to the equations in (5.11) on p. 79.
2. Obtain the tangent vector in (5.12) (p. 80) not by implicit differentiation, but rather by using parametric equations $(x(t), y(t))$ of the ellipse and computing $(x'(t), y'(t))$, evaluated at the appropriate value of t.
3. Adapt to hyperbolas our proof of the reflection property for ellipses.
4. Traditionally, the **latus-rectum** of a conic is a line segment through a focus, perpendicular to the axis through that focus, with endpoints on the conic. The **semi-latus-rectum** is half of that, going from the conic just to the axis, and its length is often denoted p.

(a) Show that for the ellipse

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}=1$$

(with $a>0$), we have $p=\frac{b^2}{a}$.

(b) What is p for the hyperbola

$$\frac{x^2}{a^2}-\frac{y^2}{b^2}=1?$$

(c) Find p for the parabola $y=x^2$.

(d) Try to make your answer in (c) a limiting case of p in (a); of p in (b).

(Note: the earliest occurrence of the term *latus rectum* seems to be in some lectures on conic sections written by Gilles Personne de Roberval (1602–1675). These notes were published posthumously in 1693.)

5. (a) On a bright cloudless day, suppose the elliptical shadow of a volleyball on the horizontal ground has eccentricity ε, traditionally defined as the ratio of the distance between the ellipse foci to the major axis length. In terms of ε, how many degrees above the horizon is the sun?

(b) The volleyball can also cast a shadow on a wall. Find the "wall eccentricity" in terms of the "ground eccentricity."

6. Calculus books often give this "curvature formula" for a twice-differentiable function $y=y(x)$:

$$\kappa(x)=\frac{y''}{(1+(y')^2)^{3/2}}.$$

(a) Use this formula to find the curvature of the ellipse

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}=1$$

at each of its vertices $(\pm a, 0)$ and $(0, \pm b)$.

(b) One of your formulas bears a close relation to the latus rectum p. What is this relation?

(c) Is there a corresponding relation in the case of a hyperbola? A parabola?

7. Find a family of ever-taller ellipses which approach, in the limit, the parabola $4ay=x^2$. (In this equation, a is the distance from the parabola's vertex to its focus.) Show that the real ellipse foci do *not* go off to infinity in this process, but instead approach the focus $(0, a)$ of the parabola.

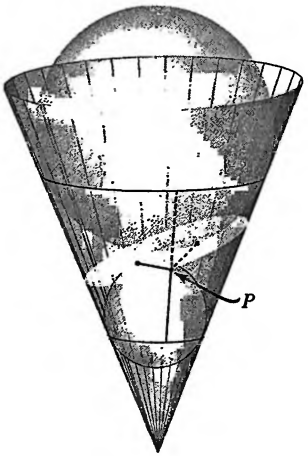
8. Let P and Q be two distinct points on a parabola with focus F, and let α be the angle formed by the segments PF and QF. Suppose the lines tangent to the parabola at P and at Q intersect at G, and let β be the angle formed by the segments PG and QG. What simple relation is there between α and β? Prove your conjecture.

9. An exercise on normals: Let the tips of the outward-directed unit normals to the ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

trace a curve. If $a = b$, this curve is a larger circle. If $a \neq b$, is it a larger ellipse?

10. In the picture below, the ellipse is a plane section of the right circular cone. Underneath, a small sphere previously dropped into the cone gradually grows until it just touches the plane of the ellipse, and above, a larger sphere shrinks until it, too, just touches that plane. Each sphere is thus tangent to both cone and plane. Remarkably, those two points of contact are precisely the ellipse foci! This discovery is attributed to the Belgian mathematicians Adolphe Quetelet (1796–1874) and Germinal-Pierre Dandelin (1794–1847). Prove this theorem. [Hint: In the picture, P is an arbitrary point of the ellipse. The two dashed segments to P are of the same length. (*Why?*) The two solidly-drawn segments to P likewise have equal length. Now use the basic stakes-and-string property.]



11. Suppose the cone in the above picture is replaced by a cylinder, the plane continuing to cut the cylinder in an ellipse. If from each end of the cylinder you drop a perfectly-fitting sphere, do the two spheres still touch the plane of the ellipse at its foci?

12. Consider the two closely-related surfaces defined by $x^2 + y^2 - z^2 = 0$ (a cone) and $x^2 + y^2 - z^2 + 1 = 0$ (a hyperboloid of two sheets). Let P be any point of the hyperboloid, and let $T(P)$ be the plane tangent to the hyperboloid at P.

(a) Prove that $T(P)$ cuts the cone in an ellipse.

(b) Prove that P is the center of this ellipse.

(c) Orthogonally project P and the ellipse to the (x, y)-plane. Is the projection of the center point the center of the projected ellipse?

(d) Formulate and establish analogues of (a)–(c) if the hyperboloid of two sheets is replaced by the hyperboloid of one sheet given by $x^2 + y^2 - z^2 - 1 = 0$.

6

Are Hyperbolas *Really* Ellipses?

in which insisting upon unity forces additional revelations upon the trio

Unifying Formulas for Foci

Teacher. Everything about foci on the x-axis applies equally well to foci on the y-axis, so we can neatly summarize our progress.

Old viewpoint: (6.1)

The ellipse

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}=1$$

has two foci. When $a \geq b$, they are located at $(\pm\sqrt{a^2-b^2}, 0)$; when $a \leq b$, they're at $(0, \pm\sqrt{b^2-a^2})$.

New viewpoint: (6.2)

The ellipse

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}=1$$

has four foci—one pair on each principal axis.

On the a-axis, the foci are $(\pm\sqrt{a^2-b^2}, 0)$; on the b-axis, they are $(0, \pm\sqrt{b^2-a^2})$.

Philosopher. Our new viewpoint is smoother and more general. But the other shoe must inevitably drop: *what about hyperbolas*? The traditional focal length of either hyperbola

$$\frac{x^2}{a^2}-\frac{y^2}{b^2}=1 \quad \text{or} \quad -\frac{x^2}{a^2}+\frac{y^2}{b^2}=1$$

is $\sqrt{a^2+b^2}$. Not only does this seem qualitatively different from what happens with ellipses, but we're also back to one pair of foci per hyperbola. My pipe dream of perfect equality among conics always seems to run into some problem!

Teacher. Maybe, but so far we've been able to break down walls, attaining greater unity. Let's keep the faith!

Philosopher. OK, I'll continue dreaming. To be truthful, though, even the very *form* of the hyperbola equation bothers me! The minus sign not only makes things look different, but invites—in fact, seemingly forces us—to split things up into separate cases. I think the equation of a hyperbola should look exactly like the equation of an ellipse. There should be total unity throughout.

Teacher. That's absolutely outrageous. It's crazy, and it is also so beautiful, it just might be suggesting a deeper truth that we're missing. But I don't know how to accommodate you on this one.

> If I have to choose between the correct and the beautiful, I'll choose the beautiful.
>
> Attributed to Hermann Weyl

> It often happens that the requirements of simplicity and beauty are the same, but where they clash the latter must take precedence.
>
> P.A.M. Dirac (*Proc. R. Soc. Edinburgh* **59**, 122, 1939)

Student. If the minus sign bothers you, why not just put it somewhere else? Why not rewrite

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1 \quad \text{as} \quad \frac{x^2}{a^2} + \left(-\frac{y^2}{b^2}\right) = 1?$$

Teacher. That doesn't really look like an ellipse equation. The second term ought to have the form $\frac{y^2}{B^2}$.

Philosopher. But I like the idea of somehow relocating the negative sign. It may look crazy, but one *could* just let B be ib which, when squared, gives us the negative sign. That is, write

$$\frac{x^2}{a^2} + \frac{y^2}{(ib)^2} = 1.$$

Student. That sure *does* look weird!

Teacher. Maybe it does, but sometimes an idea that seems off-the-wall is exactly what's needed to make a breakthrough. Isn't it true that something appears crazy if it doesn't seem to make common sense? Common sense, however, is a two-edged sword: while it often keeps us grounded, other times it keeps us *too* grounded. History is full of examples where common sense proved too parochial, too limited, based as it is on everyday experience and impressions. It can lead us away from the truth rather than toward it.

Student. Do I sense a story here?

Teacher. Why not? A few centuries ago, didn't nearly everyone believe the earth was flat? Even now, look about—it honestly doesn't look very *round*, does it? Sure, today there are lots of pictures of earth taken from spacecraft, but back then, flatness seemed pretty sensible. For that matter, it once seemed just as sensible that the earth was stationary, instead of turning on an axis in addition to flying around the sun at an even greater speed, as some astronomers were beginning to claim. In every other case of motion, when you walk, ride a horse, travel in a wheeled conveyance or ship, it's usually easy to sense the motion—visually, aurally, kinesthetically. But remain very quiet, with every sense straining, and you still feel absolutely *no motion* while riding spaceship earth! The fact is, our senses evolved for the very practical purpose of survival, not for scientific research. Physical data-gathering "senses" appropriate to that level of truth-seeking often must be so large, that we can't possibly carry them around. Early telescopes effectively "grew" our light-gathering pupils to a diameter of a couple of inches, and these have grown to hundreds of inches, weighing many tons.

> Einstein's special theory, with its twins paradox; his general theory with its warped four-space; the conundrums of quantum mechanics; an expanding universe; black holes—all were originally greeted by many as "crazy" ideas.

Philosopher. What you say emboldens me. Let's explore in what sense

$$\frac{x^2}{a^2}+\frac{y^2}{(ib)^2}=1$$

is like an ellipse. The imaginary quantity certainly tells us we're still working in the complex setting, which is where we've already had some success.

Teacher. Yes, and we've been helped by sketching a 3-D slice of the conic, as in Figure 5.4 [p. 73]. I think we should do that for a hyperbola, mimicking what we did earlier. Starting with an ordinary hyperbola

$$\frac{x^2}{a^2}-\frac{y^2}{b^2}=1$$

with real a and b, we have:

$$\frac{x^2}{a^2}-\frac{y^2}{b^2}=1$$

$$\frac{(x_1+ix_2)^2}{a^2}-\frac{(y_1+iy_2)^2}{b^2}=1$$

$$\frac{x_1^2-x_2^2}{a^2}-\frac{y_1^2-y_2^2}{b^2}+2i\left(\frac{x_1x_2}{a^2}-\frac{y_1y_2}{b^2}\right)=1$$

$$\frac{x_1^2-x_2^2}{a^2}-\frac{y_1^2-y_2^2}{b^2}=1, \qquad \frac{x_1x_2}{a^2}-\frac{y_1y_2}{b^2}=0.$$

Setting in turn x_1, x_2, y_1, y_2 equal to zero gives the four coordinate 3-D slices of $\mathbb{C}^2$. To include the original hyperbola, set x_2 or y_2 equal to zero. Choosing $y_2 = 0$, for example, yields the original hyperbola and the set defined by

$$-\frac{x_2^2}{a^2} - \frac{y_1^2}{b^2} = 1,$$

which is empty. On the other hand, choosing $x_2 = 0$ yields the original hyperbola together with the ellipse

$$\frac{x_1^2}{a^2} + \frac{y_2^2}{b^2} = 1,$$

giving a more informative picture.

Figure 6.1 shows this last situation.

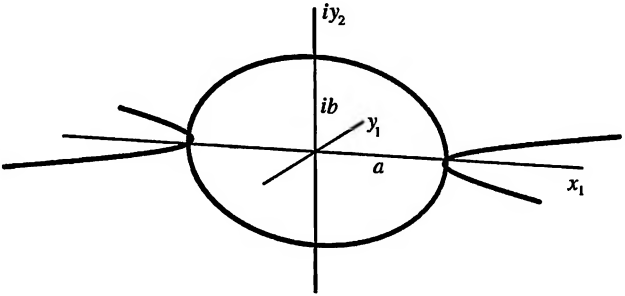


FIGURE 6.1. This picture does for a hyperbola what Figure 5.4 on p. 73 does for an ellipse. It's a 3-D slice of a complex hyperbola showing its real hyperbola branches being connected by an ellipse in the Minkowski (x_1, iy_2)-plane. Of course the ellipse is only a tiny part of the complex hyperbola, which actually turns out to be a smooth manifold of real dimension 2 in $\mathbb{R}^4$.

Philosopher. Wow! I'm looking at the ellipse in Figure 6.1, and am absolutely amazed: its semi-axes are precisely a and ib! Associated with our hyperbola is an ellipse exactly fitting the description we thought was crazy a minute ago.

Teacher. Things are actually better than that—it's not that we're looking at two curves, a hyperbola and an ellipse, but rather little parts of one larger whole which, writing

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1,$$

we can call a complex hyperbola. But writing that same equation as

$$\frac{x^2}{a^2} + \frac{y^2}{(ib)^2} = 1,$$

we could just as well call it a complex ellipse.

> An entire complex conic is like the proverbial elephant, and we are like the blind men who perceive two- or three-dimensional slices and pronounce them "ellipse" or "hyperbola"—the same way feeling the elephant's side makes one think it's a wall, while the tall is like a rope, and a leg is like a tree trunk.

Philosopher. Does this new way of writing a hyperbola solve the problems with hyperbola foci? Is there, for example, a hyperbola analogue of (6.2)?

Teacher. Well, we've succeeded in writing a hyperbola as an ellipse, so let's apply what we know about ellipse foci and see what falls out!

For an ellipse with semi-axes a and ib, (6.2) yields, for foci on the x-axis,

$$(\pm\sqrt{a^2-(ib)^2},0).$$

And for...

Philosopher (Interrupting). Wait a minute! What you just wrote is exactly the same as $(\pm\sqrt{a^2+b^2},0)$. *That hyperbola center-to-focus distance formula I was upset about turns out to be just a corresponding ellipse formula in disguise!* Perhaps we'll get the unification I wanted. For this hyperbola, let me see what we get for foci along the y-axis. (6.2) yields $(0,\pm\sqrt{(ib)^2-a^2})$.

Teacher. This becomes $(0,\pm i\sqrt{a^2+b^2})$. *It appears that we have uncovered an additional pair of foci for the hyperbola, just as we did for the ellipse!* For the hyperbola

$$\frac{x^2}{a^2}-\frac{y^2}{b^2}=1,$$

there is a focus-pair along each of the principal directions, located according to (6.2) when the equations are written in ellipse form.

Philosopher. I confidently predict analogous good news for the "vertical" hyperbola,

$$-\frac{x^2}{a^2}+\frac{y^2}{b^2}=1.$$

Teacher. Let's check! As an ellipse,

$$-\frac{x^2}{a^2}+\frac{y^2}{b^2}=1$$

has semi-axes ia and b. Applying (6.2) gives x-axis foci $(\pm i\sqrt{a^2+b^2},0)$ and y-axis foci at $(0,\pm\sqrt{a^2+b^2})$. The real answers coincide exactly with traditional formulas, but we see them in a new light. The imaginary foci are new ones we've discovered in our pursuit of beauty and unity.

> As with ellipses, for any hyperbola there is a pair of foci on each principal complex axis.

Triangles of "Vital Statistics"

Philosopher. One of my favorite facts about ordinary ellipses, and to my mind a very remarkable one, is the Pythagorean relation connecting the ellipse's two principal axis lengths and the distance between its foci: these three lengths form a right triangle. One can halve each of these three lengths and draw a picture as in Figure 6.2:

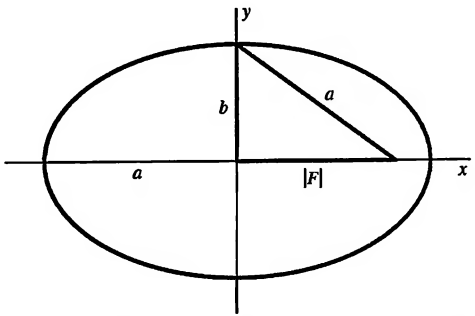


FIGURE 6.2. The traditional picture of an ellipse's fundamental right triangle. The hypotenuse length is the semi-major axis length. The triangle degenerates to a double line segment when the ellipse is a circle.

In Figure 6.2, the length of the hypotenuse is always the semi-major axis.

In the case of a hyperbola, we still have a right triangle, but the role of the sides is different: In Figure 6.3, we see that now the focal length is the triangle's hypotenuse, while for ellipses, it was one of the two triangle legs.

Pretty as these traditional pictures may be, they nonetheless do what many other traditional conic pictures do: splinter things into seemingly disparate cases. The hyperbola picture looks fundamentally different from the ellipse picture. My question now is, "How do we unify these disparate right-triangle pictures?"

Teacher. I think we should try to capitalize on what we already know about foci. Let's begin with Figure 6.2, showing an ordinary ellipse, with $a > b$. The left and right foci F_1, F_2 are the two points in which a circle of radius a, centered at the point $(0, b)$, intersects the x-axis. As a decreases toward b, these two intersection points approach one another, coalescing to a double point when $a = b$. But now this morphing process grinds to a halt—when $a < b$, the circle isn't large enough to intersect the x-axis.

This picture may well have been contemplated by the ancients. However, geometry for them was a matter of stakes, strings, straightedges and compasses—all geometric toois. They had no elementary algebra as we know it, no analytic geometry, no equation of a circle like $x^2 + y^2 = 1$. It was algebra that led mathematicians to complex numbers and permitted us to expand from the real to the complex setting. Importantly, this provides an

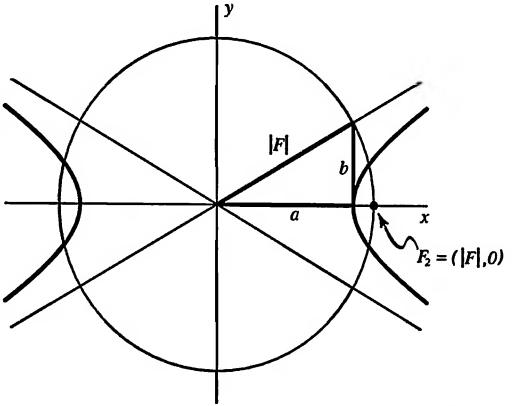


FIGURE 6.3. The traditional picture of a hyperbola's fundamental triangle. The hypotenuse is no longer the semi-major axis, but rather the focal length. How will our trio smooth out this wrinkle?

escape from the geometric straitjacket binding the ancients, because today we can translate any circle into algebraic terms through writing its equation. If we do this to our circle of radius r centered at $(0, b)$, we can, from the ancients' standpoint, "do the impossible": *we can continue the morphing process.*

Philosopher. I guess many of the traditional pictures that have caused me problems were inherited from a much earlier day. I'm very curious about how this morphing process looks as we continue it.

Teacher. We can translate into algebra both circle and x-axis: the circle is

$$x^2 + (y - b)^2 = a^2,$$

and the x-axis is $y = 0$. Setting $y = 0$ in the circle equation gives

$$x^2 = a^2 - b^2, \quad \text{or} \quad x = \pm\sqrt{a^2 - b^2},$$

valid for any choice of a and b. Starting at $a > b$, as a steadily decreases, the foci coalesce to $x = 0$ along the real part of the complex x-axis, and then they separate along the imaginary part of the x-axis. This is very similar to what we saw in Figure 5.3 [p. 71] as the complex parabola was pushed up.

Philosopher. But what about the triangle?

Teacher. We can follow that along, too. The triangle starts in the usual (x_1, y_1)-plane, as in Figure 6.2, with altitude b lying on the y_1-axis. The base lies on the x_1-axis, and has

length equal to the focal length $\sqrt{a^2 - b^2}$, which at this stage is still a real number. Suppose we look at that triangle with base extending from (0, 0) to $(+\sqrt{a^2 - b^2}, 0)$, again as in Figure 6.2. When the base length becomes 0, the triangle collapses to a double line segment from (0, 0) to $(0, b)$. As a continues to decrease, $\sqrt{a^2 - b^2}$ now becomes imaginary, so the base extends from (0, 0) to this imaginary number in the ix_2-axis. Before the triangle collapses to a line segment, it lives in an ordinary Euclidean plane. After collapse, it reemerges in (ix_2, y_1)-space, a Minkowski plane. Suppose, for example, that a has decreased to 1. Since $b = 2$, the triangle looks like this:

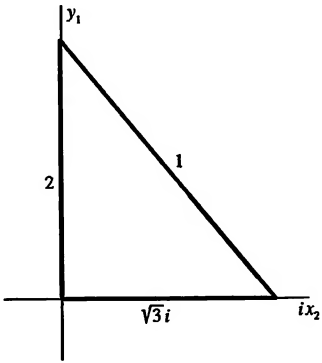


FIGURE 6.4. A "right triangle" in a Minkowski plane: $1^2 = (\sqrt{3}i)^2 + (2)^2$.

Student. You must have labeled the sides wrong—you've got the hypotenuse having length 1, while you've marked one of the legs 2. Both legs look shorter to me. Isn't the *hypotenuse* always the longest side?

Teacher. That's true in Euclidean space, but it need not be true in the Minkowski plane where our triangle lives. This becomes clear when you notice that the other leg is marked with a pure imaginary number, so the separation-squared between its endpoints is *negative*. Thus, when we compute the squared length or separation of the hypotenuse using the Pythagorean theorem, the horizontal leg, when squared, makes a negative contribution. Result: although the hypotenuse looks longer than the vertical leg to our Euclidean eyes, its length in Minkowski space is actually less, due to the negative contribution of the horizontal leg squared.

Philosopher. So there's an "x-axis triangle" for every choice of a and b, no matter which of them is larger. When $a > b$, the triangle is Euclidean, but after collapsing to a line segment, it moves into the Minkowski (ix_2, y_1)-plane as a becomes less than b.

Teacher. Yes, and there's also a "y-axis triangle" for every choice of a and b, no matter which of them is larger. When $b > a$, the triangle is Euclidean, living in the (x_1, y_1)-plane, and it moves into the Minkowski (x_1, iy_2)-plane as b becomes less than a.

Philosopher. That means that Figure 6.2, showing as it does an ellipse with $a > b$, really has *two* fundamental right triangles, one in a Euclidean plane, the other in a Minkowski plane. We can see a more complete triangle-picture by looking at a 3-D slice, like this:

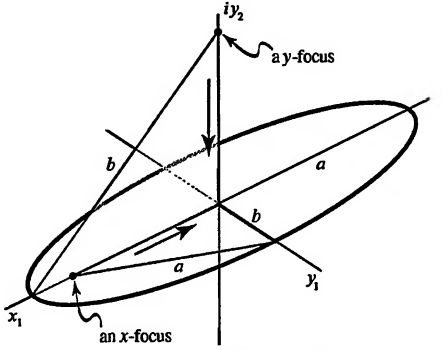


FIGURE 6.5. Extending Figure 6.2 to this 3-D slice of $\mathbb{C}^2$ gives us room to draw a fundamental triangle with respect to each principal axis. In this illustration, $a > b$; a is decreasing and b is fixed, so both the x-focal length $\sqrt{a^2 - b^2}$ and y-focal length $\sqrt{b^2 - a^2}$ are decreasing, indicated by the arrows. Both the darker Euclidean and lighter Minkowski triangles are headed for collapse...

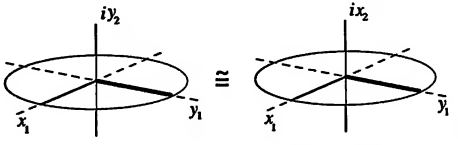


FIGURE 6.6. ...as we see here. This is the situation when $a = b$. In each picture, the heavy segment is the collapsed Euclidean triangle and the other solid segment is the collapsed Minkowski triangle. The collapsed view in the (x_1, y_1)-plane is common to both these 3-D slices, but as a continues to decrease the action shifts to the three-dimensional slice shown in the right picture, and...

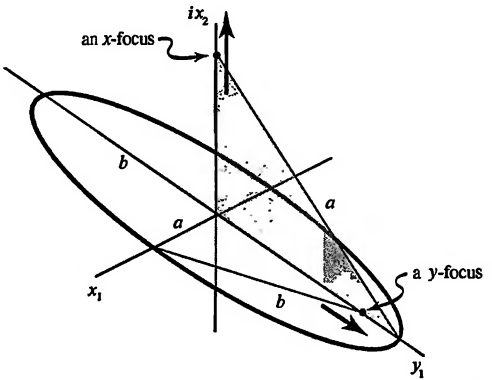


FIGURE 6.7. ...we see this situation after a has become less than b. The darker Euclidean triangle in Figure 6.5 has morphed into the darker one here, which is now Minkowski. The lighter, Minkowski triangle in Figure 6.5 has morphed into the lighter, Euclidean triangle, seen here.

> The traditional fundamental right triangle of an ellipse always has a mate. It lives in a Minkowski plane containing the ellipse's imaginary focus-pair.

Notice that since the x-focal length is $\sqrt{a^2 - b^2}$ and the y-focal length is $\sqrt{b^2 - a^2}$, at all times in the morphing process, these appear equally long. Also, note the transitional nature of Figure 6.6, connecting (x_1, y_1, iy_2)-space and (x_1, y_1, ix_2)-space.

Teacher. You assumed an "$a > b$" ellipse. Starting with a "$b > a$" ellipse, the sequence of shots would just be reversed.

Philosopher. Right.

Interactive Applets

For animated versions of Figures 6.5 and 6.7, see Chapter Applets 4 and 5 in the CD-ROM.

Chapter Summary

We saw in the last chapter that one pair of foci lies on a real axis, the other on an imaginary one. This imaginary axis turns out to be a powerful unifying idea. For example, a

hyperbola's equation

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$$

can be written as an ellipse equation:

$$\frac{x^2}{a^2} + \frac{y^2}{(ib)^2} = 1.$$

This realization leads to foci formulas that assume a single consistent form, just as standard equations for ellipses and hyperbolas now do. This leads to an additional discovery: the traditional right triangle involving semi-axes a, b and focal length has a "mate" extending last chapter's discovery of the new set of foci, and Figures 6.5–6.7 show how the two triangles work together. Chapter Applets 4 and 5 in the CD-ROM give an animated version of this.

Exercises

1. In Figure 6.5 on p. 97, we haven't shown the hyperbola (living in the Minkowski plane) attached to the ellipse. Sketch in that hyperbola, and notice where its foci are. We're used to seeing the foci of a hyperbola on the "convex side" of each branch. Here, they're not. Is this consistent? Explain.
2. Add all four foci to Figure 6.1 on p. 92. Again, is the placement of the ellipse foci consistent?
3. The pictures in Figures 6.5–6.7 illustrate triangle-pairs for

   $$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1.$$

   Sketch an analogous sequence of pictures for

   $$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1.$$

4. Here's an exercise on area measurement in the Minkowski plane: the units of length in the x-and iy-directions are 1 and i, so it is natural to assign an area of $1 \cdot i$ to the square having vertices $(0, 0)$, $(1, 0)$, $(0, i)$ and $(1, i)$. If we regard a region in the Minkowski plane as a limit of ever-finer covers by small coordinate-squares, it makes sense that the region's area is pure imaginary. Thus, for example, the lighter triangle in Figure 6.5 should have imaginary area and, in fact, the "half base × height" formula does give

   $$\text{Area} = \frac{ia\sqrt{a^2 - b^2}}{2}.$$

   As a check for consistency, see if Heron's formula for a triangle's area gives the same result. Heron's formula says that if s is the semi-perimeter $\frac{1}{2}(a + b + c)$ of a triangle with sides a, b, c, then the triangle's area is $\sqrt{s(s-a)(s-b)(s-c)}$.

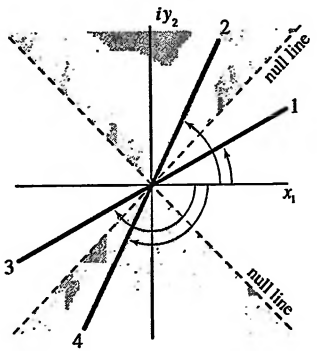
5. One can extend the notion of angle to the Minkowski plane: if $\mathbf{v} = (a_1, ia_2)$ and $\mathbf{w} = (b_1, ib_2)$ are two vectors in the Minkowski plane, then the standard relation

$$\mathbf{v} \cdot \mathbf{w} = \|\mathbf{v}\|\|\mathbf{w}\| \cos(\theta)$$

can be used to *define* the angle between v and w—namely,

$$\theta = \text{Arccos}\left(\frac{a_1 b_1 - a_2 b_2}{\sqrt{a_1^2 - a_2^2} \cdot \sqrt{b_1^2 - b_2^2}}\right)$$

Here are some computer-experiments to provide some intuition about angle-measurement in this environment. Consider this picture:



In the Minkowski plane, release a photon at an arbitrary point P (which, in the picture, we've taken to be the origin). We choose space (x_1) and time (iy_2) units so that the photon's history is represented by an upward-pointing ray on a null line. Now nothing can travel faster than a photon, so the null lines through P divide the plane into possible future, possible past, impossible future and impossible past. This gives rise to "right-left" angles terminating in an unshaded area, and "up-down" ones ending in a shaded area.

(a) Using Maple's **arccos** command, demonstrate that any right-left angle α is pure imaginary, and that it approaches $\pm i \cdot \infty$ as α approaches a null line. (To make Maple return a complex floating point number, write the full command using a decimal point as, for example, **arccos(100.0*I)**, where the decimal point is included.)

(b) Demonstrate that any up-down angle β has the form $\frac{\pi}{2} - (\textit{pure imaginary})$, and that the size of the imaginary part approaches infinity as β approaches a null line.

(c) Demonstrate, and then prove: if two lines are symmetric with respect to a null line, then the angle between them is $\frac{\pi}{2}$—that is, their dot product is 0. Show that every right angle in the Minkowski plane appears to our Euclidean eyes as symmetric about a null line. (In the picture, we've drawn things so that each odd-numbered line segment drawn from the origin makes a right angle with each even-numbered segment.)

(d) Do the angles of any triangle in the Minkowski plane sum to π?

6. The two triangles in Figure 6.5 have the same side lengths, in the sense that in each triangle, the real or imaginary parts of the real or imaginary sides are: a, b, $\sqrt{a^2 - b^2}$. (The imaginary part of any complex number is real—for example, the imaginary part of $2 + 3i$ is 3.) Thus there is some sort of "congruence" between the two triangles. If this idea has validity, then the angles should also correspond in some sense. Can you form such a correspondence?

7

Stakes and Strings

...in which—at last—
the puzzle of subtracting
lengths gets solved!

In Celebration of Ellipse Foci

Philosopher. I've been thinking again about the "stakes-and-string" construction of an ellipse. It's so beautiful the way string lengths always add up to the diameter of the ellipse! But why on earth do we have to use the *difference* of string lengths for a hyperbola? And how come we get only one branch at a time? And if we interpret a string segment as a light path, why does a photon head the wrong way after bouncing off a mirrored hyperbola? Everything works so nicely for an ellipse, but at every turn, it seems, a hyperbola has to behave differently. I wonder if there's any hope of further unifying ellipses and hyperbolas.

Teacher. Our *modus operandi* has worked pretty well so far: insist that hyperbolas and ellipses are at their core the same, and if necessary, break out of conventional thinking to arrive at a deeper unity between them. Let's begin by trying to extend our understanding about the sum-of-lengths and reflection properties of an ellipse. Then we'll attempt to stare down hyperbolas!

There's actually a basic example that combines the sum-of-lengths and reflection properties of an ellipse—a "whispering gallery." An elliptical room has two mutual "sweet points": the foci. Stand at one focus, and you can hear someone whispering at the other. Theoretically, all the sonic energy emanating from one focus goes through the other focus. It's a "focus-to-focus phenomenon."

> It's said that John Quincy Adams, as a member of the House of Representatives, often pretended to doze off at one of the focal points of what is today called "Statuary Hall," in the U.S. Capitol building. In fact, however, he was secretly listening in as others carried on hushed discussions at a table near the other focus.

Philosopher. You're making an analogy between light and sound, no? A short burst of sonic energy is like a burst of photons released when a flashbulb goes off. In either case, the surface acts as a reflector—a light-mirror or a "sonic-mirror."

Teacher. Yes. Actually, many things that bounce off an appropriate ellipsoidal surface will do this—radio waves, infrared or "heat" waves, billiard balls...

Student. It's lucky I didn't bet anything when I took a tour of the Mormon Tabernacle in Salt Lake City, Utah! The guide put me at one focus of the huge ellipsoidal room, and then he walked way over to the other focus. I did not believe him when he said I'd hear a pin drop from his focus. But I heard it clear as a bell, almost as if it were dropped right next to me!

Teacher. Actually, what you heard demonstrates more than the focus-to-focus phenomenon, which by itself doesn't ensure hearing the pin drop.

Student. What do you mean?

Teacher. This property says that from an elliptical reflecting surface, a sonic packet leaving one focus heads, after reflection, directly toward the other focus. That's one piece of ellipse magic. But there's another: the sum-of-lengths property. Although the energy packets, like billiard balls, may leave one focus in a variety of different directions, the total distance traveled in going from one focus to the other is always the same. In the Mormon Tabernacle, the first property says only that (in theory, of course) all the sonic energy of the dropping pin passes through where you're standing. The second property additionally says that packets simultaneously leaving one focus all arrive simultaneously at the other focus. That preserves the sound's wave shape. Without that, you might hear an attenuated hiss, not a pin drop.

Student. I never thought about that!

Philosopher. I know a nice way to see these two properties working in tandem. Drop a pebble directly over one focus of a very quiet ellipse-shaped swimming pool. A circular ripple expands around that focus, and after a little while, points of the ripple begin reflecting off the ellipse. When the reflecting stage is over, we see a circular ripple contracting toward the other focus. The ripple converges to a point, but because all the packets comprising the ripple continue on in straight-line paths, we now see a ripple expanding away from this second focus. After an equal time, this ripple completely reflects and converges to the original focus. There's an endless, periodic exchange of energy from one focus to the other, somewhat reminiscent of a pendulum swinging back and forth.

Teacher. Yes, but...

Student (Interrupting). If it's periodic, what is its period?

Philosopher. The circle expands at a uniform rate since all the little energy packets travel at the same speed. The total distance traveled from focus to focus is equal to the diameter of the ellipse, so if the diameter is $2a$, then in a full period each packet travels $4a$. If the packets' common speed is r, then the full period is $4a/r$.

Student. Are you seriously saying that I should have heard the pin drop over and over again, every $4a/r$ seconds?

Philosopher. No. You were standing at a focus, and when the sonic energy converged, you absorbed most of that energy. It got converted into other forms, mostly trace quantities of heat. Because so much of the energy gets absorbed after the first reflection, there's very little sonic energy left and you can't hear much of anything more. If the energy weren't absorbed in this and other ways (by the walls, for example), someone whispering at one

focus would receive his own echo, and continued whispering would build up to a deafening roar!

Student. I'll take the real world.

How Far to a Focus?

Teacher. Consider the familiar ellipse

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}=1$$

with $a > b$, and let F_1 and F_2 be its left and right foci lying on the real x-axis. Earlier, we found formulas—(5.5) and (5.6) [p. 77]—for the distance from F_i to a point (x, y) on this ellipse. The point (x, y) is this far from the left focus F_1:

$$L_1(x)=a+\frac{\sqrt{a^2-b^2}}{a}\cdot x$$

and this far from the right focus F_2:

$$L_2(x)=a-\frac{\sqrt{a^2-b^2}}{a}\cdot x,$$

showing that the two distances add to the ellipse diameter $2a$.

Student. The graphs of $y = L_1(x)$ and $y = L_2(x)$ are just straight lines!

Teacher. Yes, and we can actually plot the ellipse and these two lines all on the same piece of paper, like this:

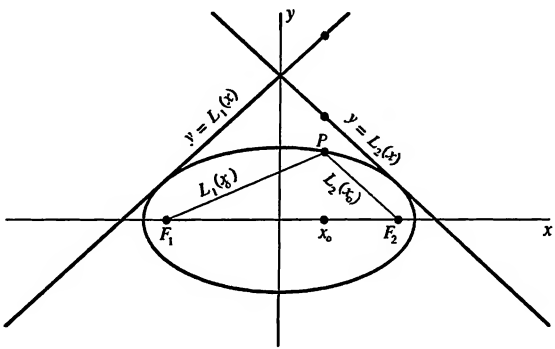


FIGURE 7.1. The distance functions from an ellipse focus to a point P on the ellipse with horizontal coordinate x_o. $L_1(x)$ is the signed distance function from the left focus F_1, and $L_2(x)$ is the signed distance function from the right focus F_2.

We read Figure 7.1 this way: a point on the major axis determines two points on the ellipse. We've sketched in one of these points, $P = (x_o, y_o)$. The length of $\overline{F_1P}$ is $L_1(x_o)$, and the length of $\overline{PF_2}$ is $L_2(x_o)$.

Interactive Applets

For interactive versions of Figure 7.1, see Chapter Applets 6 and 7 on the CD-ROM.

Student. The lines seem like they're tangent to the ellipse. Are they, actually?

Teacher. To answer that, we could begin by solving for where either line intersects the ellipse. To do this for $L_1(x)$, for example, substitute $L_1(x)$ for y in the ellipse's equation. This gives a quadratic, from which we can find x. The line will intersect in 0, 1 or 2 points, corresponding to 0, 1 or 2 distinct real solutions. A "1" outcome would correspond to a double point, which we can regard as the limit of secant lines (a secant line intersecting the ellipse in two distinct intersection points) approaching the tangent line at (x, y) (one double point).

Philosopher. So we're hoping for a repeated root!

Teacher. Yes. Go back to $L_1(x)$ [p. 105] and substitute this for y in

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1.$$

Doing that leads to

$$x^2 + 2x\sqrt{a^2 - b^2} + (a^2 - b^2) = 0,$$

which factors into $(x + \sqrt{a^2 - b^2})^2 = 0$, a perfect square! Therefore $-\sqrt{a^2 - b^2}$ is a repeated root, telling us that the graph of $L_1(x)$ is in fact tangent to the ellipse, as it appears. A similar argument shows that the same is true of $L_2(x)$.

Student. Are these points of tangency anything special?

Teacher. Well, that repeated solution $x = -\sqrt{a^2 - b^2}$ is just the left focus! So $L_1(x)$ touches the ellipse right over that focus.

Philosopher. Similarly, then, the graph of $y = L_2(x)$ touches the ellipse over the right focus?

Teacher. Yes. Each of these lines touches an end of a latus-rectum.

Student. OK, I'm wondering about something else, too. Are the points where $L_1(x)$ and $L_2(x)$ cross the x-axis anything special?

Teacher. To find out where $y = L_1(x)$ crosses the x-axis, we just set $L_1(x) = 0$ and solve for x. Ditto for $L_2(x)$. We get

$$x = \mp\frac{a^2}{\sqrt{a^2 - b^2}}$$

for the left and right points of intersection. It turns out that the vertical lines through these points are indeed special—*they're the directrices of the ellipse.*

Student. I remember hearing about directrices, but I really don't remember what they are.

Teacher. Take a line D, a point F not on the line, and a positive number ε. Let PF denote the distance from P to F, and let PD denote the distance from P to the line D. The set of all points P for which

$$\frac{PF}{PD} = \varepsilon \tag{7.1}$$

is an ellipse if $\varepsilon < 1$, a parabola if $\varepsilon = 1$, and a hyperbola if $\varepsilon > 1$. ε is known as the **eccentricity** of the conic, D is a **directrix**; and F is a focus. In the case of an ellipse or hyperbola, the other focus similarly determines a directrix, too, so there are actually *two* directrices.

For a given focus and directrix, Figure 7.2 shows how the resulting conic depends on the eccentricity:

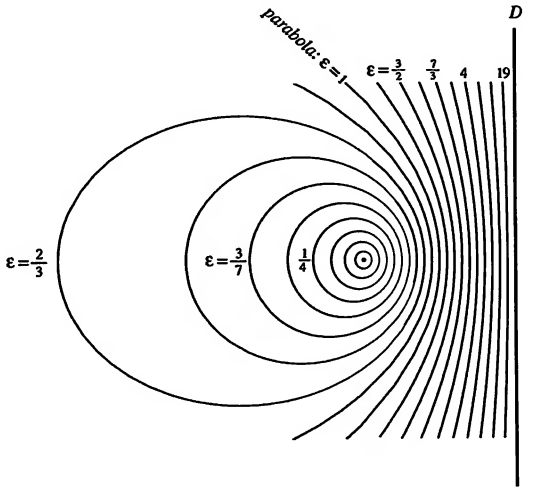


FIGURE 7.2. Conics of eccentricity $\varepsilon = \frac{r}{1-r}$, for 19 r-values: 0.05, 0.1, 0.15 ··· 0.95. If the focus (the point near the middle of the smallest, nearly-circular ellipse) is the origin, and the directrix D is $x = 1$, then r is where the conics cut the positive x-axis. Values of ε range from $\frac{1}{19}$ to 19. Notice that as $\varepsilon < 1$ increases toward 1, the ellipse becomes ever wider relative to its height.

Philosopher. This seems to be another example of where a constant defines a conic.

Teacher. I agree. "Sum-of-lengths equals a constant" is an addition (or subtraction) condition, while (7.1) is an essentially multiplication (or division) condition.

Over the Edge

Student. I'm still looking at the graphs of $y = L_1(x)$ and $y = L_2(x)$. It seems that for them to make sense, we should erase the parts below the x-axis.

Philosopher. That would put the erased points in a different category. I want to *remove* exceptions, not introduce them!

Student. But whoever heard of *negative distances*?

Philosopher. They can represent directed distances.

Teacher. I agree. The numbers of the real line can be thought of as directed distances from the origin. This idea is used all the time with functions—they take on positive or negative values. When you integrate a function like $y = -1$ from, say, 0 to 10, you get a negative area, not a positive one.

Student. But what distances are the negative numbers $L_1(x)$ and $L_2(x)$ actually measuring?

Teacher. First, just from looking at Figure 7.1, we see that when $L_1(x)$ is negative, x is to the left of where the ellipse intersects the x-axis. Similarly, when $L_2(x)$ is negative, x is to the right of where the ellipse intersects the x-axis. Now Figure 7.3 shows a 3-D slice of the complex ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1,$$

with $a > b$:

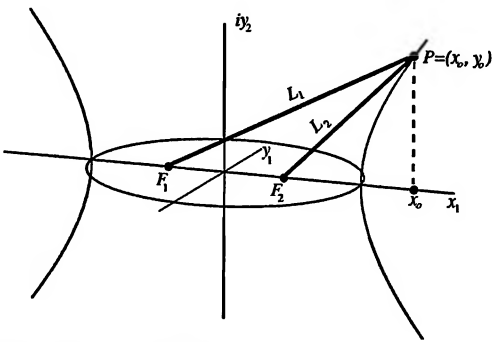


FIGURE 7.3. Here x_o has increased beyond the rightmost real point of the ellipse. As a consequence, $P = (x_o, y_o)$ has climbed up the right hyperbola branch in the Minkowski plane. But lengths $L_1$and L_2 continue to increase and decrease, respectively. always summing to $2a$.

The hyperbola branches in the Minkowski (x_1, iy_2)-plane are similar to the attached branches in Figure 5.4 [p. 73]. Figure 7.3 shows that as we increase x beyond a, we start to climb up (or down) the right branch of the hyperbola in the (x_1, iy_2)-plane.

Philosopher. Do you think, then, that $L_1(x)$ and $L_2(x)$ are measuring distances from the foci to points on the hyperbola?

Teacher. Yes. As soon as $P = (x, y)$ climbs up the hyperbola, its y-coordinate becomes pure imaginary, so the square of this coordinate is negative. This means that in Figure 7.3, even though the line segment $\overline{PF_2}$ appears to grow to our Euclidean eyes, its separation $L_2(x)$ continues to decrease. In a way, it's a continuation of a process started when $x = 0$. Then, both line segments had the same length or separation. As x increases, $\overline{F_1P}$ in effect takes string length away from $\overline{PF_2}$. *This process is only partially completed by the time x reaches a, and it continues as x further increases.* At the x where the graph of $L_2(x)$ crosses the x-axis, $L_1(x)$ has 100% of $2a$, and $L_2(x)$ has 0%.

Philosopher. This sheds a new light on something important! We saw a minute ago that $L_2(x)$ crosses the x-axis at

$$x = \frac{a^2}{\sqrt{a^2 - b^2}},$$

which is just the x-intercept of the right directrix. So it seems that the complex ellipse, whose Minkowski hyperbola branches appear in Figure 7.3, intersects the complex directrix at this special 100%–0% stage.

Teacher. Yes! And on the left side of the ellipse, the graph of $y = L_1(x)$ similarly intersects the other complex directrix at the 0%–100% stage.

Student. You mean at the 100%–0% stage, the L_2-segment has zero length? It doesn't look that way, but I guess it's the Minkowski plane again.

Teacher. That it is, and it makes for a remarkable picture:

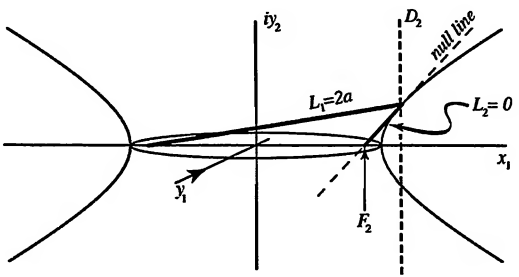


FIGURE 7.4. This depicts the remarkable situation of L_1 having the entire length of $2a$, and L_2 having none. The segment L_2 lies on a null line, on which the separation between any two points is zero.

The line segment corresponding to "$L_2 = 0$" has a very special inclination within the Minkowski (x_1, iy_2)-plane. To our Euclidean eyes, the segment appears to make a 45° angle with the x_1 -axis, but in the Minkowski plane it actually has slope i. Figure 7.5 suggests how the separation between any two points on a line of slope $\pm i$ is zero; these are null lines. They have a higher-dimensional analogue in special relativity, where they become "light cones."

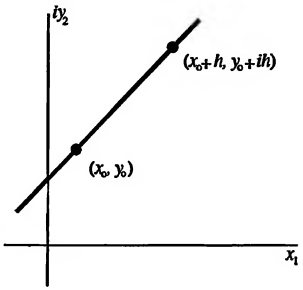


FIGURE 7.5. The point-to-point formula shows that the separation between any two points on a null line is 0:

$$\text{sep}^2 = ((x_o + h) - x_o)^2 + ((y_o + ih) - y_o)^2 = h^2 - h^2 = 0.$$

Philosopher. That seems to say that in the "right triangle" having hypotenuse length $L_2 = 0$ in Figure 7.4, the two shorter legs look to our Euclidean eyes as if they are equally long.

Teacher. Yes, and we can check that they really are. The x_1-coordinate of F_2 is $\sqrt{a^2 - b^2}$ and the directrix D_2 has x_1-coordinate

$$\frac{a^2}{\sqrt{a^2 - b^2}}.$$

The horizontal leg of the triangle is then

$$\frac{a^2}{\sqrt{a^2 - b^2}} - \sqrt{a^2 - b^2} = \frac{a^2 - (a^2 - b^2)}{\sqrt{a^2 - b^2}}$$
$$= \frac{b^2}{\sqrt{a^2 - b^2}}. \tag{7.2}$$

The vertical leg can be found by substituting

$$\frac{a^2}{\sqrt{a^2 - b^2}}$$

for x in the ellipse equation and solving for y. This leg turns out to be

$$\frac{ib^2}{\sqrt{a^2-b^2}}, \tag{7.3}$$

which is i times the horizontal leg.

> When we remove exceptions, we may at first seem to be doing something crazy, but at the end of the day, it may have led to a deeper truth.

Then What Happens?

Philosopher. Suppose one pushes x beyond the 100%–0% point—*then* what?

Teacher. The null line is tangent to the hyperbola, and Figure 7.4 shows that except at this point of tangency, all points of the hyperbola lie below the null line. This means that for all x_1, the line segment labeled L_2 is always space-like (real), never time-like (imaginary). Same for L_1. This dovetails with the formulas for $L_1(x)$ and $L_2(x)$, which are both real-valued. But after $L_1(x)$ or $L_2(x)$ reaches zero, it becomes negative. That is, when we push x beyond the 100%–0% point, $L_1(x)$ or $L_2(x)$ becomes negative.

Philosopher. Certainly the *function* $L_2(x)$ becomes negative, but what right do we have to say that the length of the line segment L_2 itself is then negative? We crafted a formula working for an ellipse, but it seems to me that now we're extending it beyond its design limits.

> Any physical theory works only within a certain range—it has a design limit. Take Newton's mechanics, for example: At too small a scale, it doesn't work, and quantum mechanics had to be developed to extend the range. Newtonian mechanics likewise fails when masses are too large, or distances or speeds too great. Einstein's theories extend these ranges. Theories simply break down when stressed too far, and in mathematics, too, pushing design limits of various formulas or theorems often leads to failure. But surprisingly often, they continue to work, revealing a deeper side. Our trio often finds itself in this happy state of affairs.

Remember, $L_1(x)$ in (5.5) [p. 77] was obtained by taking a square root of

$$L_1^2(x) = \left(a + \frac{\sqrt{a^2-b^2}}{a} \cdot x\right)^2;$$

we took the positive square root because positive numbers appropriately model ordinary distances. Same for $L_2(x)$, and I see no problem with all this. But what right do you have to choose the *negative* root, now?

Teacher. Your objections are well taken, and they deserve a good answer. Our approach to extending conics has always been to maintain the defining polynomial form and look at

complex solutions. We did this earlier, when we extended the concept of distance-squared to separation-squared, and even in finding angles. Now $L_1(x)$ and $L_2(x)$ are real, first degree polynomials, and their graphs of course dip below the x_1 -axis. Furthermore, they add up to $2a$, which is surely a nice thing. Now, suppose we did elect to consider only non-negative solutions. Then our distance functions would be $|L_1(x)|$ and $|L_2(x)|$, and these absolute value functions not only fail to be polynomials, they also don't add up to $2a$. Since we've already seen how, in calculus and analytic geometry, we deal all the time with directed distances, why not go with the flow, and let $L_1(x)$ and $L_2(x)$ do what they're trying to do?

Philosopher. You seem to be saying that my gut reaction to take positive distances is "mathematics." But "Mathematics" itself comes down on the side of maintaining the polynomial form of the distance functions.

Teacher. That's a good way to put it.

Philosopher. I'll provisionally agree with this choice. But let's see how things ultimately pan out.

Teacher. Fair enough. If the truth be told, $L_1(x)$ and $L_2(x)$ sum to $2a$ for *all* complex values of x, not just for real values along x_1. It seems to me that down the road, this opens the way for an even more-inclusive understanding of conics.

For now, the bottom line is that as x assumes arbitrary real values, $L_1(x)$ and $L_2(x)$ provide consistent results. As x_1 travels far to the right, L_1 grows without bound, continuing to "take distance away from L_2" just as it did for small values of x_1. As x_1 travels to the left, L_1 decreases without bound, always giving distance to L_2. In all cases, the distances add to $2a$, which is the property that started this whole discussion!

Are Hyperbolas Any Different?

Philosopher. At the beginning of this discussion I was wondering why one must use the *difference* of string lengths for a hyperbola. And why we get only one branch at a time! I'm hoping we can adapt our distance functions $L_1(x)$ and $L_2(x)$ to a hyperbola. If we're lucky, maybe we will finally see what's going on.

Teacher. It doesn't seem all that difficult to adapt $L_1(x)$ and $L_2(x)$ to a hyperbola—after all, we know how to look at the equation of a hyperbola as the equation of an ellipse. Because the hyperbola

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$$

can be written as the ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{(ib)^2} = 1,$$

and because deriving our functions $L_1(x)$ and $L_2(x)$ was a purely algebraic process, we can formally use a and ib in place of a and b in our formula. This means that the square-root quantity becomes

$$\sqrt{a^2 - (ib)^2} = \sqrt{a^2 + b^2},$$

so

$$L_1(x) = a + \frac{\sqrt{a^2+b^2}}{a} \cdot x; \quad L_2(x) = a - \frac{\sqrt{a^2+b^2}}{a} \cdot x.$$

Notice that these still add to $2a$. Furthermore...

Philosopher (Interrupting). Wait a minute—you just said something very profound! We're now talking about hyperbolas, and you just noticed that these *ADD* to $2a$! That's what I've wanted all along. Could solving my problem really be that easy?

Teacher. That would be nice, wouldn't it?

Philosopher. But we need to check that adding the hyperbola version of $L_1(x)$ and $L_2(x)$ coincides with subtracting lengths in the usual textbook sense. For example, if $P = (x, y)$ is a point on the right branch of the hyperbola, then we'd expect that $L_1(x)$ should be positive, while $L_2(x)$ should be negative. *Adding* these would of course be the same as *subtracting* a positive length L_2 from a positive length L_1.

Teacher. So the question is, where do $L_1(x)$ and $L_2(x)$ go negative? That's easily answered: $L_1(x) = 0$ and $L_2(x) = 0$ gives

$$x = -\frac{a^2}{\sqrt{a^2+b^2}} \quad \text{and} \quad x = \frac{a^2}{\sqrt{a^2+b^2}},$$

respectively.

Philosopher. Does this mean that for a point (x, y) on the right branch of the hyperbola, $L_1(x)$ is positive and $L_2(x)$ is negative?

Teacher. Yes, it does. Obviously (x, y) is on the right branch whenever x is to the right of a, and $L_2(x)$ is negative for all x to the right of a.

Philosopher. How come? I can understand what you're saying only if a itself were to the right of

$$\frac{a^2}{\sqrt{a^2+b^2}}.$$

Teacher. But that is automatically the case, because a being to the right of

$$\frac{a^2}{\sqrt{a^2+b^2}}$$

simply says

$$a > \frac{a^2}{\sqrt{a^2+b^2}},$$

and it's easy to see that this inequality is in fact true. Reason: $b^2 > 0$, so obviously

$$1 > \frac{a}{\sqrt{a^2+b^2}}.$$

Now multiply both sides by a. Since a is positive, this gives us what we want:

$$a > a \cdot \frac{a}{\sqrt{a^2+b^2}} = \frac{a^2}{\sqrt{a^2+b^2}}.$$

This means that $L_2(x)$ is *already negative by the time x starts climbing up the right branch*, at $x = a$. Furthermore, $L_1(x)$ is positive for

$$x > -\frac{a^2}{\sqrt{a^2+b^2}},$$

so $L_1(x)$ is positive on the right branch. Here's a picture illustrating the situation:

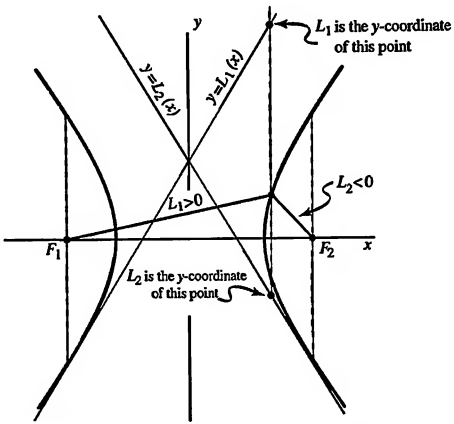


FIGURE 7.6. Here are the distance functions $L_1(x)$ and $L_2(x)$ sketched over a hyperbola. Note that at all x corresponding to the right branch, $L_1(x)$ is positive and $L_2(x)$ is negative. These signed distance functions therefore sum to $2a$, which is to say the conventional, corresponding *difference* of unsigned (positive) distances is $2a$.

Philosopher. All this is truly excellent news. It appears that an analogous argument works for the left branch.

Teacher. Yes. Everything works perfectly, because in that case, L_1 is always negative, and L_2 is always positive.

Philosopher. In Figure 7.6, those vertical dotted lines seem to be indicating that even for a hyperbola, the point of tangency lies above the foci.

Teacher. That's right, because once again all our reasoning was algebraic and carries over verbatim to hyperbolas.

Philosopher. And directrices?

Teacher. Well, we can write our hyperbola as an ellipse:

$$\frac{x^2}{a^2} + \frac{y^2}{(ib)^2} = 1,$$

so the directrices are

$$x = \pm \frac{a^2}{\sqrt{a^2 - (ib)^2}}.$$

This is, of course,

$$x = \pm \frac{a^2}{\sqrt{a^2 + b^2}},$$

so we get traditional formulas for hyperbola directrices.

Philosopher. And for a hyperbola with branches going up and down, the arguments are all essentially the same?

Teacher. Yes. The distance functions from the foci on the y-axis are

$$L_1(y) = b + \frac{\sqrt{a^2 + b^2}}{b} \cdot y; \quad L_2(y) = b - \frac{\sqrt{a^2 + b^2}}{b} \cdot y.$$

Entering the Inner Void

Philosopher. When we got the distance functions for an ellipse, it seemed remarkable that we could take x beyond the left or right vertex and measure separations of "string segments" from the foci to $P = (x, y)$. Even though the segments were in the Minkowski plane, their separations still summed to $2a$. One could similarly venture beyond hyperbola vertices. In this case, though, instead of going farther away from the origin as we did with ellipses, we now go *closer* to the origin. I'm sure everything has an analogue, but it would be nice to see a specific picture.

Teacher. I agree. Figure 7.7 shows a three-dimensional slice of $\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$.

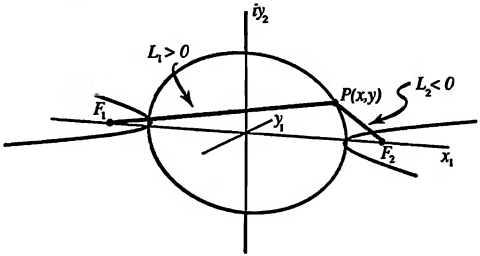


FIGURE 7.7. In this illustration, x has been chosen slightly to the left of the right branch. Compare this with Figure 7.6, noticing that $L_2(x)$ still appears to be negative. However, in the next figure...

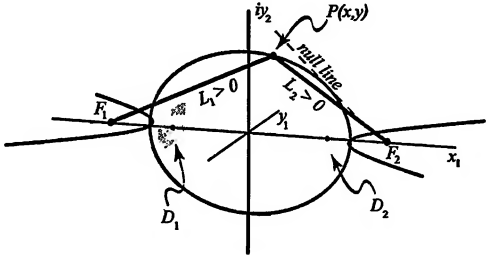


FIGURE 7.8. ...the point $P(x, y)$ on the Minkowski ellipse has been chosen with x so close to the origin that *both* signed distance functions $L_1(x)$ and $L_2(x)$ are positive. This far into the void between the hyperbola branches, the *unsigned* distances from foci to point *add* to $2a$. Interestingly, this particular 3-D slice of $\mathbb{C}^2$ contains both y_1 and iy_2, so it contains the *entire complex directrices* D_1 and D_2, suggested by the shaded patches.

Although for $|x| \geq a$ we always get one positive and one negative L_i, this doesn't hold throughout the "excluded region" $|x| < a$. In Figure 7.7, x has ventured just a little distance into the void, but not so far that L_2 has yet changed sign. Figure 7.8 is different, for here x has moved a little to the left of the directrix D_2, and here, both L_1 and L_2 are positive. This directrix passes through a point of tangency of a null line with the conic. This null line is dashed in Figure 7.8.

Figure 7.8 shows the two complex directrices

$$x_1 = \pm\frac{a^2}{\sqrt{a^2+b^2}},$$

and indicates that $L_1(x)$ and $L_2(x)$ are *both* positive for

$$|x| < \frac{a^2}{\sqrt{a^2+b^2}}.$$

As a specific example, take $x = 0$; then our L_i formulas give $L_1 = L_2 = a$. Here's a direct check, using the Pythagorean theorem to get the distance from F_1 or F_2 to $(0, ib)$:

$$L_i^2 = \left(\pm\sqrt{a^2+b^2}\right)^2 + (ib)^2 = a^2 + b^2 - b^2 = a^2.$$

Chapter Summary

As a point P travels along an ellipse from one end of the major axis to the other, the two string segments $\overline{F_1P}$ and $\overline{PF_2}$ sum to a constant, and one can think of the shortening segment as donating length to the growing one. The real picture would have us believe that this process stops when we get to the other end of the major axis, but in this chapter we discover that this give/take process continues indefinitely. Figure 7.1 on p. 105 shows

the two central actors: the two signed string-length functions measured from each of the foci. Importantly, we find that a hyperbola has analogous signed functions, and that the traditional procedure of *subtracting unsigned* string lengths for a hyperbola comes down to *adding signed* string lengths. This is because in the customary hyperbola picture, whenever one signed length is positive, the other turns out to be negative.

Exercises

1. Consider the two lines: a null line through a focus of an ellipse; a directrix of the ellipse. These lines always intersect on the complex ellipse. What happens to this intersection point as the eccentricity of the ellipse approaches 0?
2. Find formulas for the points in which the directrices of a complex ellipse

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}=1$$

intersect that ellipse.

3. Modify your answers in (2) for the hyperbola

$$\frac{x^2}{a^2}-\frac{y^2}{b^2}=1.$$

4. (a) In Figure 7.2 on p. 107, each ellipse intersects the positive x-axis at r, where

$$\varepsilon=\frac{r}{1-r}.$$

Prove that the other point of intersection with the x-axis is

$$x=\frac{r}{2r-1}.$$

(b) Figure 7.2 shows only one hyperbola branch. Is that because the focus/directrix method encapsulated in (7.1) on p. 107 yields only one branch? Or does the method actually give both branches? If so, where does the other branch cross the x-axis?

(c) Find formulas for all real ellipse and hyperbola foci in Figure 7.2.

(d) Find formulas for all real ellipse and hyperbola directrices (drawn or not) in Figure 7.2.

5. The relation in (7.1) on p. 107 in the focus-directrix definition of a conic can be rewritten as $(PF)^2=\varepsilon^2(PD)^2$, and in this form, one can translate into rectangular coordinates after choosing F and D. For example, let $F=(0,0)$ and choose for directrix the line $x=\Delta$, where $\Delta\neq 0$. Write $P=(x,y)$.

(a) Derive the equation $(1-\varepsilon^2)x^2+2\varepsilon^2x+y^2=(\Delta\varepsilon)^2$.

(b) When $\Delta=1$, this equation describes the conics in Figure 7.2. (Note that when $\varepsilon=0$ it reduces to the equation for the origin—the real circle of zero radius appearing in the figure.) The figure suggests that the conic centers move indefinitely

leftward as ε increases from 0 to 1. Complete the square and obtain a formula for the center's position. Compare that with the average of the ellipse's two foci formulas in 4(c).

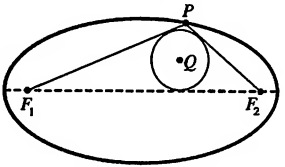
(c) Explain what happens to the center as ε increases beyond 1.

6. Philosopher notes that dropping a pebble at a focus of a quiet elliptical swimming pool creates a circular ripple that theoretically expands and contracts endlessly. Teacher was about to object to something. *What could it have been*? (Suggestion: look at the wave amplitudes on the expanding circles. Do they remain *uniform* around the circle as the scenario progresses? Try tracing the history of different small arcs of the circle.)

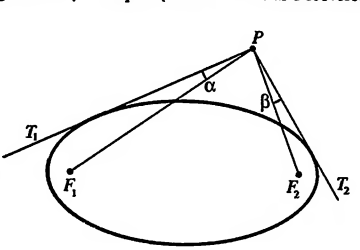
7. Even in the complex setting, one can make a statement like $L_1(x) + L_2(x) = 2a$. Apply this to stakes located at $(\pm i, 0)$, and a string of length $4i$. Sketch this "ellipse" in an appropriate plane.

8. Redo problem 7 if the string length is 4 instead of $4i$.

9. In the stakes-and-string picture below, Q is the center of the circle inscribed in triangle F_1PF_2. As P goes around the ellipse, does Q trace out an ellipse? If it does, what is its eccentricity in terms of that of the original ellipse?



10. (a) In the following diagram, P is a point outside an ellipse; T_1 and T_2 are lines through P tangent to the ellipse; F_1, F_2 are ellipse foci. Find an algebraic proof that the angles α and β are equal. (This theorem was discovered by Poncelet.)



(b) What about P *inside* the ellipse? Does your algebraic proof address this issue? Can you give a specific example to illustrate?

(c) Does Poncelet's theorem work for hyperbolas? (In the real case, assume that the foci are *inside* the topological loop in $\mathbb{P}^2(\mathbb{R})$ defined by the ellipse or hyperbola.)

11. Each curve in Figure 7.2 is the locus of points P having a constant distance ratio

$$\frac{PF}{PD}$$

from a point F and line D. By working out the appropriate algebra, create an analogue to this figure, in which points P on each curve have a constant distance ratio

$$\frac{PQ_1}{PQ_2}$$

from two points Q_1 and Q_2. Note that when the ratio is 1, the curve is a line.

8

Directrices, New and Old

in which Philosopher's insatiable thirst for symmetry leads to additional directrices

Each Focus has a Counterpart

Philosopher. I've been thinking about Figure 7.2—the one [on p. 107] showing a focus, directrix and a family of conics of various eccentricities. Although that figure doesn't show it, for each ellipse or hyperbola there's another focus as well as another directrix, creating a more symmetric setup. For an ellipse we get a picture like this:

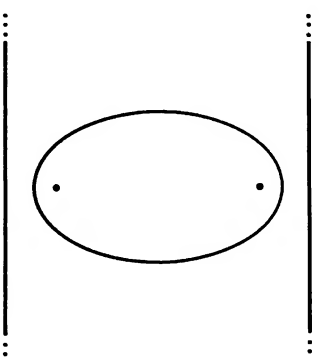


FIGURE 8.1. An ellipse with its two traditional foci and corresponding directrices. In this picture of $\frac{x^2}{5^2}+\frac{y^2}{3^2}=1$, the eccentricity ε is $\frac{4}{5}$, the distance from the ellipse's center to either focus is $\sqrt{a^2-b^2}=4$, and the distance from the center to either directrix is $\frac{a^2}{\sqrt{a^2-b^2}}=\frac{25}{4}=6.25$. Note that the geometric mean of these two distances is $\sqrt{\sqrt{a^2-b^2}\cdot\frac{a^2}{\sqrt{a^2-b^2}}}=5=a=$ the x semi-axis.

I'm willing to bet that the corresponding picture for hyperbolas looks like this:

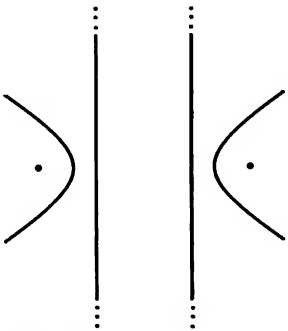


FIGURE 8.2. This is $\frac{x^2}{5^2} - \frac{y^2}{3^2} = 1$, showing the hyperbola with its traditional x-foci and directrices; its eccentricity is $\varepsilon = \frac{5}{4}$. Since $a = 5$ and $b = 3i$, we have $\sqrt{a^2 + b^2} = 4$ and $\frac{a^2}{\sqrt{a^2+b^2}} = \frac{25}{4}$. Their geometric mean is 5 which, once again, is the x semi-axis a.

Teacher. I agree. In Figure 7.2, we can successively look at a series of ellipses of increasing eccentricity. Think of it as a movie—an ellipse with its right focus and associated directrix fixed, and its left focus and directrix both moving leftward toward infinity as ε increases toward 1. When $\varepsilon = 1$ the focus has moved to the line at infinity, and the corresponding directrix is the line at infinity. The conic has become a parabola. As ε grows larger than 1, the moving focus and directrix pass beyond the line at infinity and appear on the right side of the affine plane, giving hyperbolas.

Philosopher. So for each ellipse or hyperbola, each focus has a unique mate or partner—its directrix. And conversely, to every directrix there's a focus. In Figure 8.1, the left focus and left directrix are mates, as are the right focus and right directrix. The same holds in Figure 8.2. There's a 1:1 correspondence between a focus and its directrix!

> The Latin term "*directrix*" has the spirit of "directress." It is a line which, given a focus, directs where the conic's points should be placed. "Line" is feminine in Latin, and therefore our line has the feminine "*ix*" suffix instead of the masculine "*or*," as in *director*. There are other such masculine-feminine pairs: *tractor-tractrix*; *executor-executrix*; *aviator-aviatrix*.

It would be nice to make this focus-directrix association quantitative.

Teacher. We can do that. Our job will be a little easier if we introduce some "official notation" for the distances from the conic center to the foci and directrices. Let's agree to this:

> $\|F\|$ denotes the algebraic distance $\sqrt{a^2-b^2}$ from the ellipse center to either x-focus, and $\|D\|$ denotes the algebraic distance
>
> $$\frac{a^2}{\sqrt{a^2-b^2}}$$
>
> from the ellipse center to either x-directrix. [See the box on p. 82.]

Because ellipses and hyperbolas in standard form have their center at the origin, we can regard their foci as vectors based at the origin, so that $\|F\|$ has the usual connotation of algebraic length of a vector. And if the conic *isn't* centered at the origin? I'll still write $\|F\|$ since it's so simple and suggestive—it seems to be a harmless "abuse of notation."

Philosopher. I can agree to that, but what about $\|D\|$? Is it the length of some vector, too?

Teacher. Yes. If you draw a vector from one point to the nearest point on a given line, then the length of that vector is the distance from the point to the line. In our case of conics centered at the origin, the point is the origin, and the line is the directrix D. So again, it is suggestive to write $\|D\|$ for this distance

$$\frac{a^2}{\sqrt{a^2-b^2}}.$$

Philosopher. OK, so now you're going to make our focus-directrix association quantitative with this simple notation?

Teacher. Right. Since

$$\|F\| = \sqrt{a^2-b^2} \quad \text{and} \quad \|D\| = \frac{a^2}{\sqrt{a^2-b^2}},$$

we have $\|F\| \cdot \|D\| = a^2$ This means

$$\|F\| = \frac{a^2}{\|D\|} \quad \text{and} \quad \|D\| = \frac{a^2}{\|F\|}.$$

In fact, each of $\|F\|$and$\|D\|$can be expressed individually in terms of a and ε:

$$\|F\| = a\varepsilon; \qquad \|D\| = \frac{a}{\varepsilon}. \tag{8.1}$$

Philosopher. How would you prove this?

Teacher. An easy way is to write out what (7.1) [p. 107] says for the point $P = (a, 0)$ on the ellipse. In this case, "distance to F" is $a - \sqrt{a^2-b^2}$, and "distance to D" is

$$\frac{a^2}{\sqrt{a^2-b^2}} - a.$$

After a little simplification, we find that (7.1) reduces to

$$\varepsilon = \frac{\sqrt{a^2 - b^2}}{a}.$$

This easily gives both relations in (8.1).

Interestingly, the relation $\|F\| \cdot \|D\| = a^2$ noted a moment ago can be rewritten as

$$a = \sqrt{\|F\| \cdot \|D\|}, \tag{8.2}$$

which says, geometrically, that *a is the geometric mean of $\|F\|$ and $\|D\|$*.

Student. I hate to bring up a problem, but what about a circle? We once said that its center really consists of two foci occupying the same position. Is the same true of its directrices?

Teacher. You raise a good point. Note that (8.1) implies that if we vertically stretch the ellipse in Figure 8.1—making ε decrease—the two foci draw closer together. As $\varepsilon \to 0$, the associated directrices move farther apart, approaching the line at infinity as the foci coalesce. In a circle, the foci have become coincident, and the directrices have both become the line at infinity. In a beautifully symmetric way, when either of these pairs becomes double, so does the other. In this sense the correspondence between focus and directrix continues to hold for a circle.

Philosopher. Those equations in (8.1) are very nice, and seem to give my correspondence conjecture an explicit form:

$$(a\varepsilon, 0) \leftrightarrow \left(\frac{a}{\varepsilon}, 0\right) \quad \text{and} \quad (-a\varepsilon, 0) \leftrightarrow \left(-\frac{a}{\varepsilon}, 0\right).$$

Student. That doesn't make any sense to me. When ε is 1, we have a parabola, and a parabola goes on forever. It's not symmetric about the origin.

Philosopher. But...but...I feel that my idea is quite natural, and somehow it has to be right.

Teacher. Maybe it is. Let's try slightly modifying my proposed animation of Figure 7.2. Starting with any ellipse there, increase ε toward 1. If the right focus remains stationary, then the ellipse center moves left, and the ellipse gets both larger and relatively more horizontally elongated.

Philosopher. I see that.

Teacher. To keep the changing picture looking like Figure 8.1, let's do two things: first, translate each ellipse so its center stays at the origin. Second, uniformly shrink or expand the picture so the horizontal vertices stay at, say, $(\pm 1, 0)$. Viewing this spectacle as ε increases toward 1 produces a changing ellipse with vertices fixed at $(\pm 1, 0)$ and becoming ever thinner in the y-direction. Its x-foci, which are of course inside the thinning ellipse, are symmetrically separating, approaching $(\pm 1, 0)$. Likewise, the corresponding directrices—outside the thinning ellipse—are symmetrically approaching $(\pm 1, 0)$. In the limit, when $\varepsilon = 1$, the ellipse has degenerated to the double line segment from $(-1, 0)$ to $(1,0)$. The foci are its endpoints, and the directrices are the lines $x = \pm 1$, as in this illustration:

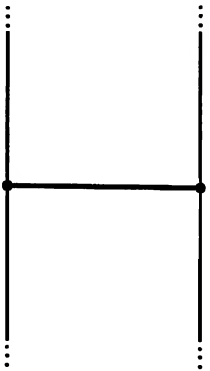


FIGURE 8.3. Not everyone has seen a parabola viewed symmetrically—here's one such view. A bounded conic of eccentricity 1 is a double, closed line segment, its foci are its endpoints, and its directrices are the two perpendicular lines through these endpoints.

So we've attained symmetry, but paid a price—the parabola has become degenerate.

Philosopher. But the correspondence still holds, in spite of the degeneracy.

Student. In your moving picture, suppose that we keep the conic centered at the origin, but keep the y-axis vertices fixed, instead? Then the ellipses couldn't collapse to a line segment. That way, perhaps we could avoid degeneracy.

Teacher. Actually, we just get a different *type* of degeneracy: the ellipse becomes two distinct horizontal lines. Both the foci and directrices race off to infinity, finally meeting there. One could regard that as one focus of multiplicity 2, together with the line at infinity as a directrix of multiplicity 2.

Philosopher Drops a Bombshell

Philosopher. Now, about my correspondence conjecture—it has an important connection to an idea that's been brewing in me...

Student. I love it when you say things like that!

Philosopher. Earlier, we found that both ellipses and hyperbolas possess a second pair of foci—there's a pair along each principal axis. *By my correspondence, shouldn't there be a second pair of directrices, too?* Just as there are foci both on the x-axis and the y-axis, independent of whether $a > b$ or $a < b$ (one pair being real, the other imaginary), there ought to be corresponding directrices defined in all cases, too, one pair being real, the other being imaginary. Figure 8.1 shows two vertical directrices, and if we stretch the ellipse vertically, its directrices stay vertical until the ellipse becomes taller than it is wide. Then the directrices apparently become horizontal. This sounds more than a little reminiscent of

the "foci-ruse"—that great hoax—that we discovered earlier. In Figure 8.1, the foci on the x-axis coalesce as we stretch the ellipse vertically, and then they seem to separate along the y-axis. But we know that actually, the foci on the real x-axis coalesce and then separate along the *imaginary* part of the x-axis, still remaining "x-foci." Simultaneously, the y-foci approach each other, but are imaginary and normally unseen. Only as they separate after meeting do they become real and hence visible in the traditional picture. Thus y-foci appear in the customary real-plane view, just as the x-foci disappear!

Teacher. So you're thinking that an analogous thing might happen with directrices?

Philosopher. I'd bet money on it! And there should be an eccentricity relative to each principal axis, too—an x-eccentricity and a y-eccentricity.

Teacher. This is amazing!

Philosopher. Pictures like Figures 8.1 or 8.2 should have three-dimensional analogues to those showing the imaginary foci on a minor axis. In these analogues, we ought to see the new directrices.

Teacher (Recovering from shock). Just like that, you come up with a whole new idea—you're painting a whole new vision! What you say sounds beautiful, and I feel you're probably right about it. *How come I didn't think of this?*

Philosopher. We now suspect two new directrices, but we need to *find* them!

Student. And draw some great pictures!

> After Dirac predicted the existence of the positron (the antimatter dual to the electron), it didn't take long to find it. Its telltale signs had been seen before, but always ignored.

Teacher. I think the approach we've been using will continue to work: maintain the algebraic form ordinarily used only in the real setting, and then see what we get as we morph the complex ellipse.

Philosopher. Everyone recognizes the canonical picture showing the x-foci of an ellipse when the ellipse is longer in the x-direction, as in Figure 8.1. So we should stretch this vertically and see what the directrix formulas give us.

Teacher. Exactly. For example, let's explore the "$a < b$" ellipse

$$\frac{x^2}{4^2} + \frac{y^2}{5^2} = 1. \tag{8.3}$$

Now the directrices in the $a > b$ case have equations

$$x = \pm \frac{a^2}{\sqrt{a^2 - b^2}} \tag{8.4}$$

and our *modus operandi* tells us to keep this equation when $a < b$. Now $a = 4$ and $b = 5$ in (8.3), so (8.4) becomes

$$x = \pm \frac{16}{\sqrt{-9}} = \mp \frac{16i}{3}. \tag{8.5}$$

The situation is illustrated in Figure 8.4 where, to keep the picture uncluttered, I've sketched in just one of the new directrices:

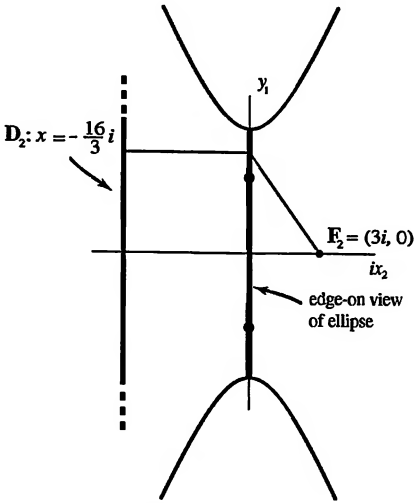


FIGURE 8.4. This picture is essentially Figure 5.7 on p. 74, but viewed looking straight down the x_1-axis. The heavy vertical line segment to the right is an edge-on view of the ellipse

$$\frac{x^2}{4^2} + \frac{y^2}{5^2} = 1.$$

This figure shows one x-focus and its associated directrix. The geometric mean of the x-coordinates of F_2 and D_2 in (8.2) is

$$\sqrt{\left(\frac{-16i}{3}\right) \cdot 3i} = 4,$$

so even when the foci are on the ix_2 axis, the x semi-axis $a = 4$ is still the geometric mean of the coordinates defining an x-focus and its associated directrix. The two heavy dots in the edge-on view of the ellipse are the real y-foci.

Philosopher. This picture is illuminating! I'd felt a little uncomfortable about Figure 7.2 [p. 107] that shows a family of conics with various eccentricities, because all the ellipses there are elongated in the direction *perpendicular* to the directrix, and that seemed pretty arbitrary. The "dual" to this phenomenon would be elongation *parallel* to the directrix, but the ellipses clearly aren't stretched that way, and I felt trapped in a corner about that. Voila!

The dual occurs in Figure 8.4, and I can rotate this a bit to get this:

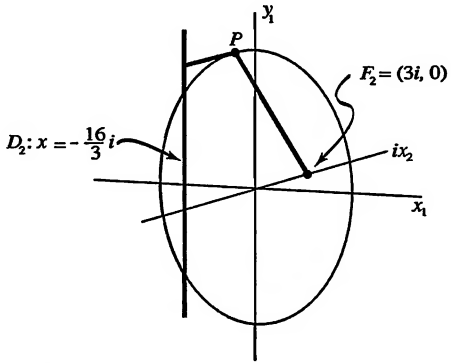


FIGURE 8.5. Here's another view of the setup in Figure 8.4—we've rotated the (x_1, ix_2)-plane about the y_1-axis. Because $\|F\| = \sqrt{a^2 - b^2}$ is imaginary and a is real, the x-eccentricity $\varepsilon = \frac{\|F\|}{a}$ is imaginary.

In each picture, the ellipse is in fact a circle stretched parallel to the directrix, and now the underlying truth seems so natural: when ε is real, elongation is perpendicular; when ε is imaginary, elongation is parallel.

Student. But aren't Figures 8.4 and 8.5 wrong? You have the focus and directrix on opposite sides of the origin. In Figure 8.1, the left focus corresponds to the left directrix, and the right focus with the right directrix.

Teacher. In (8.5), notice what the math is doing! There, i appears in the denominator, and when we multiply numerator and denominator by i, we get $\mp$, not $\pm$. As a check, let $L_2(x)$ be the distance from the focus $F_2 = (3i, 0)$ to any point $P = (x, y)$ on the ellipse. We can use either of the equations in (8.1) [p. 123] for the eccentricity. This "x-eccentricity" is easily seen to be

$$\frac{\sqrt{a^2 - b^2}}{a},$$

which in our example is $\frac{3i}{4}$. If that line D_2 is really a directrix worthy of the name, we should have $L_2(x) = \varepsilon \cdot PD_2$.

Let's now do the computation. Write P parametrically as

$$P = (x, y) = (4\cos t, 5\sin t).$$

Now (5.6) [p. 77] is

$$L_2(x) = a - \frac{\sqrt{a^2 - b^2}}{a} \cdot x,$$

which in this example becomes $4 - 3i\cos t$. And PD_2 is just the x-coordinate defining D_2 minus the x-coordinate of P—that is,

$$PD_2 = -\frac{16i}{3} - 4\cos t.$$

Thus our acid test $L_2(x) \stackrel{?}{=} \varepsilon \cdot PD_2$ becomes:

$$4 - 3i\cos t \stackrel{?}{=} \frac{3i}{4} \cdot \left(-\frac{16i}{3} - 4\cos t\right) = 4 - 3i\cos t.$$

Success!

Philosopher. This is encouraging, indeed.

> One often sees $x = a\cos t$, $y = b\sin t$ parametrizing a circle or ellipse. The assumption, of course, is that t is real. But what if the parameter isn't real? Both sin and cos can take complex arguments; what do we get? When the parameter is pure imaginary, say iu, one has
>
> $$x = a\cos iu = a\cosh u, \quad y = b\sin iu = bi\sinh u,$$
>
> parametrizing in a Minkowski plane a curve looking to our Euclidean eyes like a hyperbola. It can be shown that, more generally, as $t + iu$ runs through all complex values, the parametrization sweeps out the full complex conic
>
> $$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 \text{ in } \mathbb{C}^2.$$
>
> This is true for any $a, b \in \mathbb{C}\backslash\{0\}$.

Student. I'm still wondering about the way the focus and directrix are on opposite sides of the origin. I see that it works in our example, but would we definitely get a wrong answer by using the focus that's on the *same* side as the directrix?

Teacher. To use that focus we'd carry over (5.5) [p. 77] to the present context, getting

$$L_1(x) = a + \frac{\sqrt{a^2 - b^2}}{a} \cdot x$$

for the distance between P and the focus F_1. In our case this is $L_1(x) = 4 + 3i\cos t$, and is clearly not $\varepsilon \cdot d(P) = 4 - 3i\cos t$.

Philosopher. To pin things down, we need to generalize the above example to an algebraic proof that the directrix always works as it should, independent of whether $a > b$ or $a < b$.

Teacher. Yes, but that seems to be a straightforward rewrite. We can continue working with F_2 and D_2, since the case of F_1 and D_1 is entirely analogous. At the general level,

$$P = (a\cos t, b\sin t);$$

$$F_2 = (\sqrt{a^2 - b^2}, 0);$$

$$D_2 : x = \frac{a^2}{\sqrt{a^2 - b^2}};$$

$$\varepsilon = \frac{\sqrt{a^2 - b^2}}{a};$$

$$L_2(a\cos t) = a - \frac{\sqrt{a^2 - b^2}}{a} a\cos t = a - \sqrt{a^2 - b^2}\cos t;$$

$$PD_2 = \frac{a^2}{\sqrt{a^2 - b^2}} - a\cos t.$$

Then the relation in question, $L_2(x) \stackrel{?}{=} \varepsilon \cdot PD_2$, reduces to the identity

$$a - \sqrt{a^2 - b^2}\cos t = a - \sqrt{a^2 - b^2}\cos t.$$

Philosopher. I assume the analogous identity for the other focus and directrix is

$$a + \sqrt{a^2 - b^2}\cos t = a + \sqrt{a^2 - b^2}\cos t.$$

Teacher. Yes. And the analogous argument for the y-directrices ultimately leads to the identities $b \pm \sqrt{b^2 - a^2}\sin t = b \pm \sqrt{b^2 - a^2}\sin t$.

Student. I'm thinking back to a problem [Exercise 5 in Chapter 7 on p. 118] where we got a conic's rectangular equation using the focus-directrix definition. That's purely algebraic, isn't it?

Teacher. It sure is, and that means you can use the separation-squares in $[PF]^2 = \varepsilon^2[PD]^2$ to select points in $\mathbb{C}^2$. Thus, the full complex directrix determines all the points of a complex conic with given eccentricity ε.

Two More Pictures

Student. I'm looking at Figure 7.8 [p. 116], and I see those little patches representing *complex* directrices. Are we actually assuming complex directrices here?

Teacher. Yes. An equation like $x = \frac{-16i}{3}$ in Figure 8.4 or 8.5 naturally defines a complex line.

Philosopher. It would seem artificial to restrict it to something smaller.

Student. Well, maybe I need some more pictures. Figures 8.4 and 8.5 don't show any patches the way Figure 7.8 does.

Teacher. Free dimensions get claimed very quickly! There isn't room in Figure 8.4 or 8.5 to show the whole complex directrix, but here are two pictures that shed some light on what's going on:

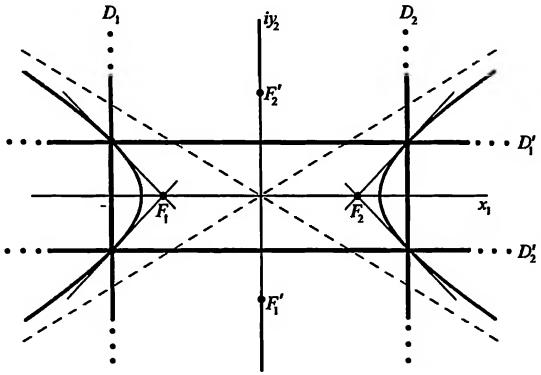


FIGURE 8.6. In this Minkowski plane, we're able to see parts of all four complex directrices of the ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 \ (a > b).$$

Extending the Minkowski plane in the y_1-direction makes the real ellipse and the complex x-directrices visible: these directrices contain the two vertical directrix traces appearing here. By extending this Minkowski plane in the ix_2-direction instead, we see the two complex y-directrices containing the two horizontal traces appearing here.

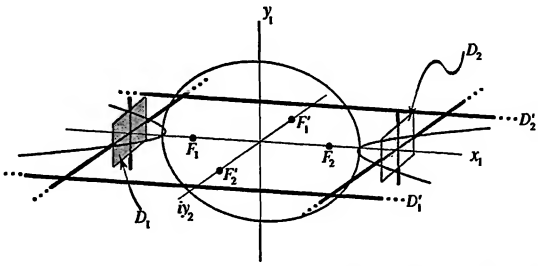


FIGURE 8.7. This illustration extends both Figure 8.1 and Figure 8.6. The new directrices are labeled D_1' and D_2' and are associated with the imaginary foci F_1' and F_2'. The real parts of these directrices are parallel to the x_1-axis. In Figure 8.1 they would be above and below the plane of the paper, parallel to the major axis.

On seeing in four dimensions

A normal human baby makes a wide variety of sounds, including ones not found in any language to which it is exposed. But through time, only its native-language sounds are reinforced, and the baby eventually loses the easy ability to make many unreinforced sounds. When we are young, our conception of space is similarly shaped and reinforced by constant exposure to our three-dimensional world. What if we don't have this reinforcement? Hassler Whitney (1907–1989), to whom this book is dedicated, was a pioneer in combinatorial and differential topology, and he once asked: can a sphere be turned inside-out in four-space, never permitting tears or creases? The fancy term for this process is "eversion," and Whitney's sphere eversion problem was eventually solved. However, the solution was highly algebraic and seemed to lack any geometric interpretation. A remarkable, step-by-step geometric eversion was found by mathematician Bernard Moran, who was totally blind from birth.

Life After Infinity

Student. One thing about these new directrices gives me a problem. When we start with an ellipse longer in the x-direction and stretch it vertically to make it circular, the x-foci come together along the real x_1-axis and then as we continue stretching, they separate along the imaginary ix_2-axis. I get that. When we do the same thing to the ellipse but instead watch the *directrices*, something escapes me. It seems that as the ellipse gets more nearly circular, $\mathcal{E}$ gets small, so according to (8.1) [p. 123], the directrices fly away to infinity!

Teacher. That's exactly what they do.

Student. What happens after that? Is there "life after infinity"? After the ellipse has become circular and then gets stretched vertically even more, what happens to the directrices? I'd like to see a movie of what happens.

Teacher. In a movie, just before the ellipse becomes circular we see two directrices symmetrically moving toward the point at infinity along the x_1-axis. After that, when the ellipse is taller than it is wide, we see two directrices traveling on the ix_2-axis, headed toward the origin.

Student. But one minute the directrices orthogonally intersect the x_1-axis of the complex plane, and then suddenly they orthogonally intersect the ix_2-axis. What does a directrix do—swing around real quick from the real to the imaginary axis?

Teacher. No, not at all, but now I see what you're concerned about! Part of the problem is the terminology "complex plane." Complex *line* is a better term, because in completing $\mathbb{C}$, only a single point at infinity is added, just as you add a single point in completing the real line $\mathbb{R}$. Thus, completing $\mathbb{C}$ is very different from completing $\mathbb{R}^2$, which acquires a whole line of points at infinity. Also, in moving toward the point at infinity, the complex line is similar to the real line in that the farther away from the origin you are, the "closer"

you are to that point at infinity. The real line, together with its point at infinity, and a set of neighborhoods about that point, defines a topology on a circle, which amounts to an ordinary circle with the usual notion of "nearness." Analogously, the complex line, together with its point at infinity, and a set of neighborhoods about that point, defines a topology on a sphere—which, similarly, is just an ordinary sphere with its usual notion of nearness. In the context of complex variables, this sphere is called the **Riemann sphere**. In the stereographic projection in Figure 8.8, look at what happens as a point in the horizontal complex line moves away from the origin—its image on the sphere gets ever closer to the sphere's point at infinity. The point in the horizontal plane could move along the x_1-axis, or it could move along the ix_2-axis. Either way, it approaches the "north pole" of the Riemann sphere.

Figure 8.8 illustrates these ideas. For simplicity, we've given the sphere a diameter of 1.

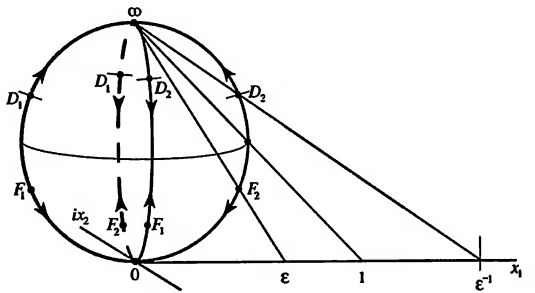


FIGURE 8.8. For a conic with vertices $(\pm 1, 0)$, let ε^2 vary in $\mathbb{R}$, from positive to negative, giving shapes similar to vertically stretching an "$a > b$" ellipse. As ε^2 varies like this, ε moves along the x_1-axis toward 0, then away from 0 along the ix_2-axis. All the while, directrices (suggested by the small whiskers) play a role reciprocal to that of the foci. For example, as foci approach 0 along x_1 and then separate along ix_2, the corresponding directrices approach ∞ along x_1 and then separate along ix_2. Points labeled F_2 and D_2 are equally far from the sphere's equator, and stereographically project to ε and $\frac{1}{\varepsilon}$ on the x-axis.

Assuming conics to have vertices at $(\pm 1, 0)$ means that $a = 1$, so (8.1) [p. 123] becomes $F = \varepsilon$ and $D = \frac{1}{\varepsilon}$, and the illustration shows this on the x_1-axis, when $\varepsilon^2 > 0$. F_1 and F_2 are below the sphere's equator, corresponding to $\varepsilon < 1$—an ellipse longer in the x-direction than in the y-direction. (Above the equator, $\varepsilon > 1$ and we have hyperbolas.) The arrows indicate that the foci are approaching the origin, meaning that the ellipse is becoming more nearly circular. According to $D = \frac{1}{\varepsilon}$, the corresponding directrices—suggested by the little whiskers—are moving toward infinity. When they've coalesced there at the north pole, the double whisker is vertical. It then separates, D_1 and D_2 traveling symmetrically away from the north pole along the ix_2-axis—the vertical great circle seen almost head-on.

Student. It seems that instead of working with the entire directrices, if you look only at the point where a directrix intersects the sphere, you get an excellent correspondence between focal points and such "directrix points." So I guess I'm not confused any more—foci and directrix points do the same thing, but at different places on the sphere.

Philosopher. It appears from Figure 8.8 that F_2 and point D_2 are equally far from the sphere's equator. A proof doesn't immediately leap out at me, but if true, that would be quite nice, with F_2 and D_2 determining, as they do, reciprocals on the horizontal x_1-axis.

Teacher. Your intuition is right, and a look at real x in Figure 8.8 makes it easy to fill in a proof:

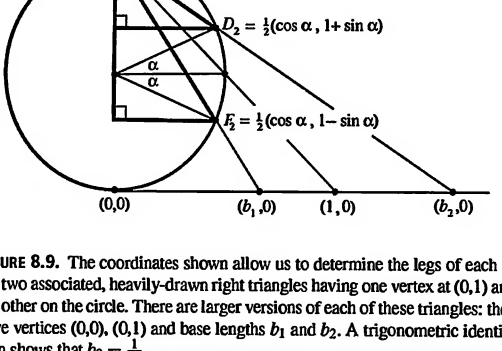


FIGURE 8.9. The coordinates shown allow us to determine the legs of each of the two associated, heavily-drawn right triangles having one vertex at (0,1) and the other on the circle. There are larger versions of each of these triangles: they have vertices (0,0), (0,1) and base lengths b_1 and b_2. A trigonometric identity then shows that $b_2 = \frac{1}{b_1}$.

Proof that in Figure 8.9, F_2 and point D_2 are equally far from the sphere's equator. Points D_2 and F_2 of Figure 8.9 are equidistant from the equator if they subtend a common angle α from the center of the sphere. It's easy to check that the three points on the right semicircle in Figure 8.9 have the indicated coordinates. The question then is: in Figure 8.9, is b_1 equal to $\frac{1}{b_2}$? In other words, is $b_1 \cdot b_2 = 1$? Working with similar right triangles in Figure 8.9 gives

$$\frac{b_1}{1} = \frac{\frac{\cos\alpha}{2}}{1-\left(\frac{1-\sin\alpha}{2}\right)} = \frac{\cos\alpha}{1+\sin\alpha};$$

$$\frac{b_2}{1} = \frac{\frac{\cos\alpha}{2}}{1-\left(\frac{1+\sin\alpha}{2}\right)} = \frac{\cos\alpha}{1-\sin\alpha}.$$

Therefore,

$$b_1 \cdot b_2 = \frac{\cos^2\alpha}{1-\sin^2\alpha} \equiv 1.$$

A Symmetric Family of Conics

Student. We introduced the Riemann sphere illustration—Figure 8.8—when we were considering conics through the vertices $(\pm 1, 0)$. I would like to see just what this family looks like. Is there a picture something like Figure 7.2 [p. 107], but with the conics symmetric about the origin?

Teacher. We should be able to get such a picture by finding an equation for its family of conics. In the spirit of Figure 7.2, let ε parametrize these conics. The standard form

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

gives us the symmetry we want, and if the conics all have $(\pm 1, 0)$ as vertices, then $a = 1$. The question then becomes one of finding b^2 in terms of ε, and either of our basic relations in (8.1) [p. 123] will do the job. For example the first one, with $a = 1$, just says $\|F\| = \varepsilon$, and we know that $\|F\| = \sqrt{a^2 - b^2}$. Therefore $1 - b^2 = \varepsilon^2$, so $b^2 = 1 - \varepsilon^2$, and our family of conics is

$$x^2 + \frac{y^2}{1-\varepsilon^2} = 1. \tag{8.6}$$

Applying a bit of graphics technology produces Figure 8.10 [on the next page]. The shaded area corresponds to negative ε^2.

Allow me to play "guide" and outline the overall idea in this illustration. We can select any one of the four moving points F_1, F_2, D_1, D_2 in Figure 8.8 and follow the conics of Figure 8.10 as that point traverses a loop. Select point F_2, for example. If we put it at ∞, the corresponding conic consists of two vertical lines through $(\pm 1, 0)$. F_1 is also at ∞, and D_1, D_2 are at 0. The point at infinity is a double focus, and the y-axis serves as a double directrix. As F_2 moves down along the right half of the circle in Figure 8.8—a real longitude of the Riemann sphere—and toward the equator, it determines the hyperbolas to the left and right of the picture's vertical strip. When F_2 is exactly on the equator, so are the other three moving points, and the foci lie on the directrices $x = \pm 1$, as in Figure 8.3 [p. 125]. The x-axis is the union of a collapsed ellipse and a collapsed hyperbola. As F_2 proceeds toward the origin, the collapsed ellipse opens up, becoming ever more nearly circular. As F_2 reaches the origin, so does F_1, giving a double focus, and D_1, D_2 both reach infinity, so they become one double directrix. Here $\varepsilon = 0$; of course (8.6) corroborates that we have a circle.

Now F_2 begins its journey up a semicircle that appears nearly head-on in Figure 8.8. The eccentricity ε is imaginary and defines the ellipses in the shaded region, all taller than wide. As F_2 nears the north pole to complete the loop, these vertically-stretched ellipses approach the two vertical lines we met at the start of our trip.

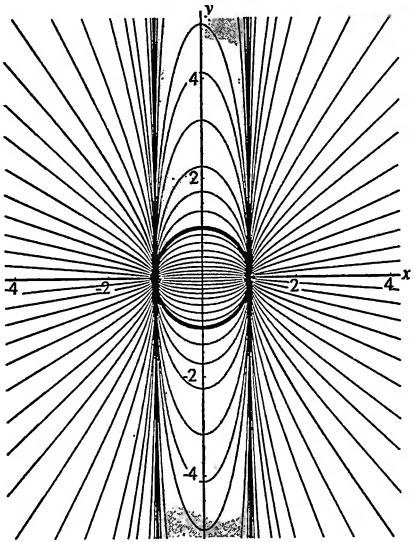


FIGURE 8.10. These are the conics with vertices $(\pm 1, 0)$ for ε^2 ranging through all real values. The hyperbolas correspond to $\varepsilon > 1$, and as ε decreases toward 1 these hyperbolas hug the x-axis ever more closely. When $\varepsilon = 1$, the hyperbolas have collapsed to two double rays and a collapsed ellipse now also appears as a double line segment between $(-1, 0)$ and $(1,0)$, giving altogether the double x-axis. As ε decreases further from 1 toward 0, the double line segues into thin ellipses in the picture's unshaded disk. These ellipses get taller, approaching the bounding circle, and they continue to grow still taller as ε^2 becomes negative—that is, as ε becomes imaginary. These ellipses, taller than wide, appear in the shaded portion.

...Just Can't Resist

Teacher. I can't resist telling you about another family of conics. This third family complements the other two we've just met. From the standpoint of magnification, the first family, shown in Figure 7.2 [p. 107], is the "smallest"—the circle is just a point. The second family, in Figure 8.10, represents an intermediate magnification, in that the circle is of finite size. The third family is the largest—the circle is infinitely large, and the family covers the plane twice. In the first family, one focus and its associated directrix are both held fixed. In the second family, two vertices are held fixed. In the third family, two foci are held fixed. This last family is parametrized by $\varepsilon^2 > 0$, and its equation is easy to derive. First, let's assume that the fixed foci are $(\pm 1, 0)$. That means that $a^2 - b^2 = 1$, or $b^2 = a^2 - 1$. Also, $a\varepsilon = \|F\|$ becomes $a\varepsilon = 1$, so $a = \frac{1}{\varepsilon}$, and the equation of these confocal conics—conics

all sharing the same pair of foci—is

$$\frac{x^2}{\left(\frac{1}{\varepsilon^2}\right)} + \frac{y^2}{\left(\frac{1-\varepsilon^2}{\varepsilon^2}\right)} = 1. \tag{8.7}$$

For $0 < \varepsilon^2 < 1$, both denominators are positive, giving nondegenerate ellipses. These approach a collapsed ellipse as ε approaches 1 from below, and a collapsed hyperbola if ε approaches 1 from above. When $\varepsilon^2 > 1$ we get nondegenerate hyperbolas. Here's the picture:

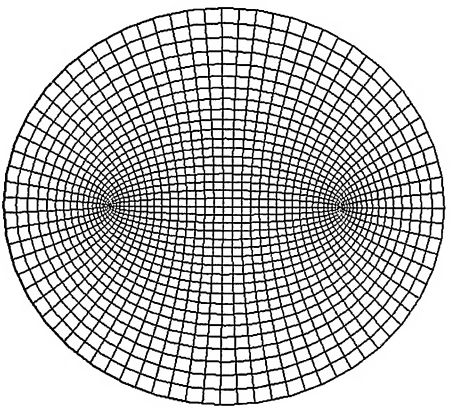


FIGURE 8.11. This turns out to be the image under $x \to \cos x$ of a Cartesian grid within a rectangle in $\mathbb{C}$—in this case, a rectangle 2π horizontally by $\frac{2\pi}{5}$ vertically. Each of the horizontal line segments in the rectangular grid maps to an ellipse, and each vertical segment maps to a hyperbola branch. Cosine, as a function mapping $\mathbb{C}$ onto $\mathbb{C}$, is "conformal"—that is, it preserves angles. This implies that the ellipses and hyperbolas intersect orthogonally in the complex y-plane since the horizontal and vertical line segments intersect orthogonally in the complex x-plane. The picture is also the image under $x \to \sin x$. Rotate the picture 90° so the foci lie on the iy_2-axis, and you have the image under $x \to \cosh x$ of the correspondingly rotated rectangle in complex x-space. The same is true of $x \to \sinh x$.

As ε varies through all real values, the family covers the plane twice—once with ellipses ($0 < \varepsilon^2 < 1$), and a second time with hyperbolas ($\varepsilon^2 > 1$). The bridge connecting these is $\varepsilon^2 = 1$, which corresponds to the line containing the two fixed foci.

Philosopher. Figure 8.11 sure looks familiar...

Teacher. Yes, this family is the image of the ordinary rectangular coordinate system under the map $y = \cos x$. The horizontal lines of the (x_1, ix_2)-plane map to ellipses, the image point orbiting an ellipse once (clockwise) for every advance by 2π along a horizontal line. When the horizontal line is the x_1-axis, the ellipse has collapsed to a double line segment, and the image similarly "orbits" back and forth along that segment. Each vertical line in the strip $-\pi < \mathrm{Re}(x) < 0$ maps to a left hyperbola branch; vertical lines in $0 < \mathrm{Re}(x) < \pi$ map to right branches. The lines $\mathrm{Re}(x) = 0$ and $\mathrm{Re}(x) = \pi$ map to right- and left-pointing rays based at $y = -1, +1$, respectively. In each instance, the usual real view—from which we get a lot of everyday intuition—corresponds to certain rare degenerate cases!

It should be mentioned that $\sin x$ behaves like $\cos x$, in that every ellipse intersects every hyperbola orthogonally. Also, note that the horizontal axes of the ellipses are never quite equal in length to the vertical axes, but become relatively more so as $\varepsilon \to 0$. In that sense, the circle has "infinite radius."

Philosopher. Figure 8.11 shows only curves corresponding to real eccentricity. What about $\varepsilon^2 < 0$?

Teacher. In that case, both denominators in (8.7) are negative—the axes are both imaginary. There is a family in the "anti-Euclidean" (ix_2, iy_2)-plane, where distances are the negatives of ordinary Euclidean distances. It is not hard to check that one gets a confocal family of ellipses, but the ellipses are now taller in the iy_2-direction than in the ix_2-direction.

Chapter Summary

There's a directrix associated to each real focus of a nondegenerate conic, so since we've found a second pair of foci, does that analogously mean there's a second pair of directrices? The answer is *yes*, and this new set removes a seemingly arbitrary phenomenon: in the usual real picture, the major axis of an ellipse is always *perpendicular* to the directrix, as if the directrix is somehow "repelling" the points of a loop to form an ellipse. The new eccentricity is imaginary, and the major axis is now *parallel* to the new directrix, as if it is now attracting the points on a loop to make the ellipse.

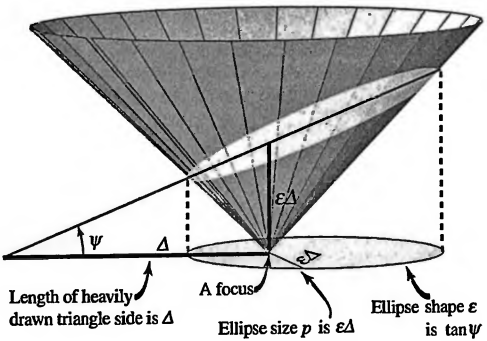
Also, just as a complex conic has two real dimensions, so also any full directrix—new or old—is a complex line and therefore has two real dimensions. In using a focus and directrix to define a conic of eccentricity ε, the conic consists of those points P in $\mathbb{R}^2$ for which $\frac{PF}{PD} = \varepsilon$. Exercise 5 in Chapter 7 (on pp. 117–118) shows that we can extend this idea to the complex setting, since there we convert the focus/directrix definition into a purely algebraic form. Therefore the full complex directrix determines the points of a complex conic of eccentricity ε.

Exercises

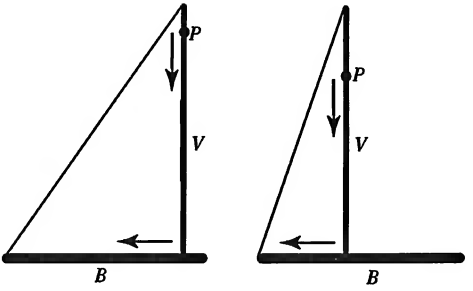
1. Verify the y-axis analogue of (8.1) on p. 123 for the ellipse

$$\left(\frac{x}{2}\right)^2 + \left(\frac{y}{1}\right)^2 = 1.$$

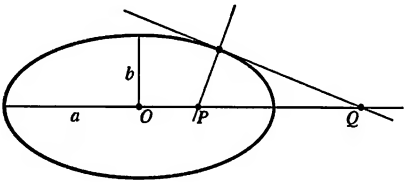
2. (a) Sketch, in an appropriate plane, the conic centered at the origin with eccentricity $\varepsilon = \frac{1}{2}$ and directrix D: $x = i$;
   (b) Redo (a) using $\varepsilon = 2$ and directrix D: $x = i$.
3. (a) In Figure 8.1 on p. 121, the point $y = 1$ on the right directrix determines two points on the ellipse. Find the distance from this directrix to the nearer point P; find the distance from P to the right focus F_2. Show that the appropriate ratio is in fact $\varepsilon = 0.8$.
   (b) Now take the point $y = 5$ on the right directrix. Carry out analogous computations (you'll encounter complex quantities) and verify that an appropriate ratio is still $\varepsilon = 0.8$.
4. (a) Extend Figure 8.5 on p. 128 by adding to it the y-axis foci and (complex)y-directrices;
   (b) Choose a not-too-special complex point on one of your y-directrices, and do the exercise analogous to 3(b).
5. (a) Can one obtain the analogue of Figure 8.11 for $\varepsilon^2 < 0$ by simply rotating that picture by 90°? Why or why not?
   (b) Prove directly that if an ellipse and a hyperbola share the same foci, they intersect at right angles.
6. Use what you've learned about conics and directrices to verify that all the labels are correct in this version of Figure A1.2:



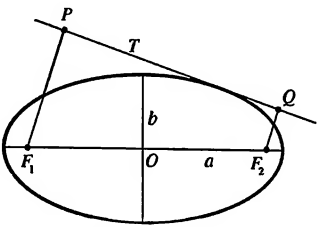
7. In the mechanical device below, base B is fixed. As suggested by the two pictures, the vertical side V can slide along B, always remaining vertical. The hypotenuse is a string attached to the base's left end; the string goes over the top end of V and the free end P of the string travels downward within V. If you keep the string taut, what kind of curve is traced out by the point $P = (x, y)$? Suppose that when P makes it down to the floor B, then P is at the origin; in this case, find the equation of the curve in terms of B, V, x, y.



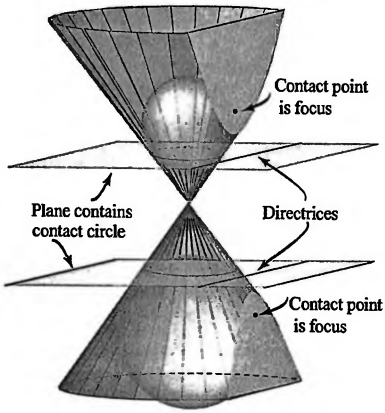
8. In this picture, the tangent and normal lines to an ellipse intersect the major axis in P and Q (assuming that neither line is horizontal). Prove that $|OP| \cdot |OQ| = a^2 - b^2$.



9. In this picture, T is a tangent line to the ellipse, and $|F_1P|$, $|F_2P|$are the distances from the foci to T. Prove that $|F_1P| \cdot |F_2P| = b^2$. Notice that this is obvious when T is horizontal.



10. The Quetelet–Dandelin theorem introduced in Exercise 10, Chapter 5 (p. 86), also holds for a hyperbola. This picture gives the idea:



State and prove the hyperbola theorem. Also, show that the directrices are the intersection of the hyperbola's plane with the planes of the two "contact circles" in which the spheres touch the cone.

11. In the "ice cream cone" sketch for Exercise 10 in Chapter 5, each circle of contact with a sphere lies in a plane, and each of these two planes intersects the plane of the ellipse in a line. Show that these two lines are the directrices of the ellipse, creating a situation similar to Figure 8.1 on p. 121.

12. In many books, the "stereographic projection" is defined as in Figure 8.8 on p. 133, but with the horizontal plane passing through the sphere's equator rather than through the south pole. Let a line through the sphere's north pole define the natural mapping from one plane to the other, and find equations describing this mapping. Does this mapping preserve angles in the two planes?

13. Consider the surfaces defined by $x^2+y^2-z^2=0$ (a cone) and $x^2+y^2-z^2+1=0$ (a hyperboloid of two sheets). Let (α, β) be any point in the (x, y)-plane. There are two points of the hyperboloid lying over (α, β); denote either one of them by P. From Exercise 12 in Chapter 5 on p. 86, we see that the plane $T(P)$ tangent to the hyperboloid at P intersects the cone in an ellipse, and that the orthogonally projected ellipse in the (x, y)-plane has center (α, β).

(a) Prove that the semi-major axis has length $\sqrt{\alpha^2+\beta^2+1}$, and that the semi-minor axis has length 1.

(b) Prove that the ellipse has eccentricity

$$\varepsilon = \sqrt{\frac{\alpha^2 + \beta^2}{\alpha^2 + \beta^2 + 1}}.$$

(c) Prove that the origin is a focus, and that a directrix for this ellipse has equation $\alpha x + \beta y + 1 = 0$.

9

Conics in General Position

in which the trio explores the two-way street connecting general conics with their equations

An "Action Reaction Principle"

Student. I've noticed that the standard form equation of an ellipse or hyperbola always puts the conic's center at the origin. Also, the axes are always horizontal and vertical. But suppose you take a wire ellipse and drop it on a Cartesian plane? It almost surely won't land in such a special way! My question is, *What's the equation of the ellipse—or of a hyperbola—after it lands in an arbitrary position?*

Teacher. A lot of the progress we've made so far has depended on maintaining an algebraic form, and the standard form simplified this. But you raise a really basic question.

Philosopher. It makes me think about my Hawaiian Islands analogy. One of Descartes' great victories was connecting geometry with algebra, and in fact it's so good, there's a "dictionary" between geometry and algebra. For example, every circle has an algebraic counterpart: its equation. Conversely, you can sketch the locus of all sorts of equations, and get corresponding geometric figures. So I agree—there *is* a big gap in what we've done so far: in the universe of conics, almost none of them have standard form equations.

Teacher. It turns out that there are three simple geometric operations that allow one to morph from perhaps the most plain-vanilla of all ellipses—a unit circle centered at the origin—into any ellipse at all. These geometric operations have equally simple algebraic counterparts that give us the corresponding equations. So we can go from the geometrically very special to the much more general by easy steps, and follow along algebraically, arriving at the equation of the wire conic. The three geometric operations are:

(i) translation in the x- or y- direction;

(ii) nonzero scaling about the origin in the x- or y-direction;

(iii) rotation about the origin.

Student. I understand translation and rotation, but not scaling. What, exactly, do you mean by that?

Teacher. Nonzero scaling in the x-direction simply scalar multiplies the x-axis by some nonzero constant. In the y-direction, it scalar multiplies the y-axis by some nonzero constant—in general, a different constant.

Philosopher. So in symbols, scaling sends (x, y) to $(\alpha x, \beta y)$?

Teacher. That's it. Now if α is positive, then the x-axis is stretched. Of course, if the positive α is less than 1, the stretch is actually a contraction, but...

Philosopher. ...but we don't want to splinter terminology! The general word "stretch" is fine with me.

Teacher. And if α is negative, that negative sign flips the x-axis about the origin—a "stretch plus flip." We have analogous statements for β. Now in our discussion of ellipses and hyperbolas, often that "flipping" part isn't important, because of symmetry. Consider a circle centered at the origin. Then using $-\alpha$ gives us the same result as α; same for $-\beta$ and β. If it doesn't make any difference this way, let's just use the simple term "stretch" instead of "scale."

Student. You can apply these transformations to the unit circle $x^2 + y^2 = 1$ to get an arbitrary ellipse? What's the idea?

Teacher. Well, first, stretch the circle horizontally and vertically to morph the circle into the ellipse shape and size you want. Next, rotate that ellipse about the origin to get the desired inclination. Finally, translate the ellipse to where you want it.

Philosopher. That's the geometry. But if I start with the circle's equation $x^2 + y^2 = 1$, exactly what do we do to this equation to attain those geometric changes (i), (ii) and (iii)?

Teacher. Let's start with translation. To translate rightward by an amount h, replace every occurrence of x by $(x - h)$, and to translate upward by k, replace every occurrence of y by $(y - k)$. Of course, moving to the right by a negative h is actually a leftward move, and going up by a negative k amounts to translating downward.

Student. But it seems to me that if you want to translate by h or k, you'd *add* by h or k, not subtract.

Teacher. Just look at an example, say $x = 0$, which defines the origin of the x-axis. To push it to the right by three units, replace x by $(x - 3)$, which gives $x - 3 = 0$, or $x = 3$. To push it to the left three units, replace x by $(x-(-3))$, giving $x = -3$. Of course in the (x, y)-plane $x = 0$ defines the y-axis, and in (x, y, z)-space, $x = 0$ defines the (y, z)-plane. In each case the object is translated.

Philosopher. Aha! Sort of like Newton's law of action and reaction—whatever we do to a variable, the variable has to *undo* to keep the equation balanced. In physics, all forces balance, algebraically adding to zero.

Teacher. That seems to be an excellent analogy, and I'll bet your idea applies more generally.

Philosopher. Wow—a general "action-reaction principle" for figures! Do you think it would work to stretch figures?

Teacher. I believe so ...to stretch in the x-direction by a factor a, such an action-reaction principle would tell us to replace every occurrence of x by $(\frac{x}{a})$, and to stretch in the y-direction by b, replace every occurrence of y by $(\frac{y}{b})$.

Student. So does that really work? Can you show me an example?

Teacher. Take a very simple example, the equation $x^2 = 1$, which defines the two points ± 1 on the x-axis. Suppose we want to stretch about the origin so the points separate to ± 2. By dividing each occurrence of x by 2, the value of x has to *double to offset that change*. That is, $(\frac{x}{2})^2 = 1$ implies $x^2 = 4$, giving $x = \pm 2$.

Student. Great! Now, about getting my general ellipse...

Teacher. Let's start by stretching the circle $x^2 + y^2 = 1$ by the factor a in the x-direction, and by b in the y-direction. Our recipe gives:

$$\left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 = 1.$$

Philosopher. Golly—this puts a whole new light on the standard form of an ellipse! Its very form expresses that the ellipse is a circle that's been stretched. I must have written those denominators a thousand times, never realizing I was always writing stretch factors!

Teacher. It's not hard to see that this stretching idea has great generality. Take an ordinary sine curve, $y = \sin x$, for example. You can stretch it both horizontally and vertically using this method, getting $\frac{y}{b} = \sin\frac{x}{a}$ —that is, $y = b\sin\frac{x}{a}$. In common parlance, the period has been multiplied by a, and the amplitude, by b. Or take the line $y = x$, stretch it vertically by m, resulting in $\frac{y}{m} = x$. Then translate up by b, which gives

$$\frac{(y-b)}{m} = x.$$

This simplifies to none other than $y = ma + b$.

Student. This is fine, but our ellipse might be *tilted*, too.

Teacher. Right, so we need to rotate. Now, to turn through an angle θ, let's apply our action-reaction principle to rotation, replacing the vector (x, y) in the equation, with wherever it is that (x, y) goes under rotation through $-\theta$. That means replacing (x, y) by

$$(x, y)\begin{pmatrix}\cos(-\theta) & \sin(-\theta)\\ -\sin(-\theta) & \cos(-\theta)\end{pmatrix}.$$

Multiplying this out gives

$$(x\cos\theta + y\sin\theta, \quad -x\sin\theta + y\cos\theta),$$

so our recipe reads: in a curve's equation, say $f(x, y) = 0$, replace (x, y) by

$$(x\cos\theta + y\sin\theta, \quad -x\sin\theta + y\cos\theta).$$

In the case of

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1,$$

we obtain, after a little rewriting:

$$\left(\frac{\cos^2\theta}{a^2} + \frac{\sin^2\theta}{b^2}\right)x^2 + \left(\frac{1}{a^2} - \frac{1}{b^2}\right)\sin(2\theta)xy + \left(\frac{\cos^2\theta}{b^2} + \frac{\sin^2\theta}{a^2}\right)y^2 = 1. \quad (9.1)$$

Notice how this reduces to

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}=1$$

when θ is 0 or π, and to

$$\frac{x^2}{b^2}+\frac{y^2}{a^2}=1$$

when θ is $\pm\frac{\pi}{2}$.

Philosopher. This is beautiful! So if $a \neq b$, we have a mixed term unless $\sin 2\theta = 0$. That is, unless θ is a whole number of right angles.

Teacher. Yes. A mixed term always implies that the conic is tilted—its principal axes are not parallel to the coordinate axes.

> A mixed term always implies tilt, but changing B in Bxy doesn't always mean the tilt changes, too. For example as $B \neq 0$ varies, the conic
>
> $$x^2+Bxy+y^2=1$$
>
> always has a 45° tilt.

Student. Is it true that to get the equation of, say, a 2-by-1 ellipse tilted 30°, all I do is substitute $a=2$, $b=1$ and $\theta = 30°$ into (9.1)?

Teacher. Yes. When simplified, that substitution gives

$$\frac{7}{16}x^2-\frac{3\sqrt{3}}{8}xy+\frac{13}{16}y^2=1. \tag{9.2}$$

Philosopher. I'm beginning to see what you mean by *general*, here—you could rotate a parabola, a cubic, a sine curve. The list is endless.

Teacher. That's right. Suppose you have a figure given by some equation, and you apply any invertible linear transformation to that figure. If that linear transformation is given by $(x, y) \to (x, y)A$ for some invertible matrix A, then to get the new equation, we replace (x, y) by $(x, y) \cdot A^{-1}$. For example, in $f(x, y) = 0$, instead of replacing (x, y) by

$$(x, y)\begin{pmatrix}\frac{1}{a} & 0\\ 0 & \frac{1}{b}\end{pmatrix}$$

—pure stretches in the x- and y-direction—we could replace (x, y) by

$$(x, y)\begin{pmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{pmatrix}\begin{pmatrix}\frac{1}{a} & 0\\ 0 & \frac{1}{b}\end{pmatrix},$$

which combines stretching followed by rotation by θ into one operation, just another invertible transformation. In this example, we get (9.1) again.

Student. We haven't said anything about the center of the ellipse. To put that at (h, k), you still have to replace x by $(x - h)$ and y by $(y - k)$ in (9.1)...

Teacher. Yes. Notice that doing this doesn't change any degree-2 term, but it does usually change linear terms as well as the constant term. The general form of a tilted conic with arbitrary center turns out to be

$$Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0. \tag{9.3}$$

Philosopher. I guess, symmetrically, one could just as easily go backwards—start with (9.3) and figure out the center and tilt of the conic.

Teacher. Whoa! When it occurs, symmetry can be wonderful and illuminating, but sometimes mathematics is decidedly unsymmetric.

Philosopher. What do you mean?

Teacher. Allow me to play "philosopher" for a minute. Algorithms in mathematics usually have an intrinsic *forward direction*, just as time has. Time moves ahead, and except possibly at quantum levels, it can't go backward. That's pretty unsymmetric! Even the everyday algorithms for adding and multiplying numbers have a directionality that makes going in reverse harder. This can vary from being only a little harder, to being so difficult that it's theoretically impossible. For example, subtracting numbers is a little harder than adding them; most people take longer to subtract two 5-digit numbers than to add them. The difference between multiplying and dividing integers is somewhat more pronounced. One usually gets unending decimals, for example.

Student. *Usually?*

Teacher. Yes. If N and M are nonzero integers, the only time $\frac{N}{M}$ *doesn't* give unending decimals is when the quotient, written in reduced form, has denominator of the special form 2^m5^n. In general, in base b, you get unending repetition unless the denominator of the reduced fraction is some product of the prime factors of b. In base ten, of course, those prime factors are 2 and 5.

Student. I never realized that about division in base b.

Teacher. And reversing multiplication in the sense of factoring a number into primes? That can get *really* difficult if the number is large. In fact, its extreme difficulty is the basis for some widely-used encryption methods, where the encryptor makes going in reverse—finding the prime factors of a number—so hard that even the fastest computer requires an impossibly long time.

Philosopher. I see your point. There really can be a fantastic lack of symmetry.

Teacher. And how easy it is to square or cube a small integer! But its corresponding roots are usually irrational—their decimal expansions continue indefinitely, without ever becoming periodic.

Philosopher. I've noticed that in the usual algorithm for computing $\sqrt{2}$, for example, the farther out you go, the harder it is to get the next digit. Sort of like stretching a spring—as you continue to stretch, the amount of work needed to stretch another millimeter increases the further out you go.

Teacher. Now when it comes to factoring polynomials, notice how quickly complexity seems to increase: there's an extremely simple formula for the root of the equation $y = ax + b$—namely, $x = \frac{-b}{a}$; the formula for a quadratic is of course fancier. Finding the roots of a third- or fourth-degree polynomial equation? For that, most people need to consult a handbook; but for its exact roots, there are nonetheless formulas that are "closed," in the sense of being a finite combination of the polynomial's coefficients that involve only $+$, $-$, $\times$, $\div$ and root-extraction. However, when you get to a general polynomial equation of degree *five*, the complexity has escalated to the point where there's no closed formula—again, in the sense of a finite combination of coefficients involving only $+$, $-$, $\times$, $\div$ and root-extraction—that gives the general equation's exact roots. And finding its roots is equivalent to finding the polynomial's linear factors.

Philosopher. I thought Newton's successive-tangent method supplied a simple algorithm for getting the zeros for all sorts of differentiable functions! And surely polynomials are differentiable.

Teacher. Yes, but that algorithm yields only a succession of better approximations to the exact zero. In addition, it is only for equations in which the coefficients have been assigned actual real (or complex) values. Asking for a closed-form formula for exact zeros of polynomials with letter-coefficients? That's much more demanding!

Philosopher. So although perhaps time can be reversed at the quantum scale, it is probably impossible at everyday time intervals such as a few days or even one second. You seem to be saying that something similar holds in mathematics—for relatively "small" problems you can reverse, but not for larger ones.

> Directionality occurs in calculus, too—differentiation is forward and integration is more like swimming upstream. Just as addition is taught before subtraction, and multiplication before division, in a beginning calculus course differentiation is usually covered before integration. In even a comprehensive handbook, the list of differentiation rules is rather short, but the list of integration formulas is neither short nor at all complete. Historically, a lot of effort has been invested in integration theory. This directionality translates to the real world: it's typically much easier to find ordinary or partial differential equations that model some physical situation than it is to solve them!

Teacher. Directionality seems to run throughout mathematics!

Now from a very general standpoint, "solving equations" is a form of reversing. Actually, there really aren't that many formulas for solving equations—linear methods and the quadratic formula come to mind as the principal ones. Fortunately with conics, though, it turns out we can reverse the process without undue difficulty. Given a conic defined algebraically by an equation of the form in (9.3), we can find its shape, size, tilt and location.

Student. The basic problem is to determine the figure from the general quadratic form in (9.3)?

Teacher. Yes, but let's focus on this specific problem:

Assume that $A, \ldots, F$ in (9.3) are real, and that (9.3) defines an ellipse or hyperbola.

Determine the shape, size and location of the conic.

We can solve this problem *exactly*. That is, the coefficients $A, \ldots, F$ defining the algebraic object in (9.3) can be transformed into *geometric* data using a finite sequence of elementary operations. In fact, five independent "parameters" are used in specifying a conic either algebraically or geometrically. For example, when none of the six letters $A, \ldots, F$ has been preassigned, we can formally divide the equation in (9.3) by any one of them, and we then have five degrees of freedom. On the geometric side one can choose shape, size, tilt, and location independently. Of course, specifying location using a conic's center (h, k) requires two additional numbers h and k, making five pieces of data in all. This is reminiscent of transforming the two coefficients of a general monic quadratic equation $x^2 + bx + c = 0$ into a "picture"—into two pieces of geometric data, the location of the two roots on the real number line. The quadratic formula represents a finite sequence of elementary operations of $+, -, \times, \div, \sqrt{}$ giving these two locations exactly.

Student. How do we go about making the transition from (9.3) to a picture?

Teacher. As we saw a minute ago, going *from* geometric data *to* the algebraic form of (9.3) is a forward process; our mission now is to reverse it. This forward-reverse process is suggestive of the computer term "LIFO"—Last In, First Out—when one is adding to or removing data from a memory stack. For us, the last forward step in arriving at (9.3) was shifting the center, and "unshifting" will be our first step in the reverse process.

Unshifting

Teacher. Let's phrase our basic unshifting problem this way: starting with an ellipse or hyperbola in $\mathbb{R}^2$ defined by

$$Ax^2 + Bxy + Cy^2 + Dx + Ey = 1, \tag{9.4}$$

find h, k and G so that when we make the substitution $x \to (x - h)$, $y \to (y - k)$, (9.4) ends up as

$$A(x-h)^2 + B(x-h)(y-k) + C(y-k)^2 = G. \tag{9.5}$$

Student. Why "G"? You've chosen the right side of (9.4) to be 1, so why wouldn't it be 1 in (9.5), too?

Teacher. Look at the example $x^2 - 2x = 1$. To complete the square, we add 1 to each side. Therefore in going from $x^2 - 2x = 1$ to $(x-1)^2 = 2$, the constant on the right side of the equation has changed.

Now let's try to massage (9.4) into (9.5). After expanding (9.5), we see that the degree-one part on the left side is

$$-2Axh - Bxk - Byh - 2Cyk. \tag{9.6}$$

To make this the degree-one part of (9.4)'s left side, equate coefficients of x, and also equate coefficients of y. This gives, respectively, these two linear equations in h and k:

$$-2Ah - Bk = D;$$

$$-Bh - 2Ck = E.$$

Now solve for h and k. The solutions can be written as

$$h = \frac{2CD - BE}{B^2 - 4AC}; \quad k = \frac{2AE - BD}{B^2 - 4AC}. \tag{9.7}$$

Philosopher. But suppose the denominator $B^2 - 4AC$ happens to be zero? Then your "solutions" are undefined, meaning there *aren't* solutions!

Teacher. You're absolutely right, and it means that sometimes it's impossible to remove both linear terms. For example, think of the parabola $y = x^2$. Its locus is actually the graph of a *function*, and that's due to the linear term y. Now the denominator $B^2 - 4AC$ is called the **discriminant** of the quadratic form $Ax^2 + Bxy + Cy^2$, and for this parabola, the discriminant is zero. On the other hand,

$$Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0$$

defines every possible conic, including parabolas as well as degenerate conics. It turns out that if the conic is an ellipse or hyperbola, which we will now assume, then $B^2 - 4AC$ is nonzero, so the equations in (9.7) really do represent solutions.

Philosopher. OK, but you still need to get G.

Teacher. I agree, but this is straightforward. With h and k as in (9.7), we compute G to be

$$G = 1 + \frac{BDE - AE^2 - CD^2}{B^2 - 4AC}. \tag{9.8}$$

Student. I generally see what you're doing, but wouldn't it be a lot easier to find the center (h, k) by just completing the square, as you did when you transformed $x^2 - 2x = 1$ into $(x - 1)^2 = 2$?

Teacher. The clinker is the xy-term. If that mixed term weren't there, then you could complete squares.

Student. I don't follow. Is there an example that shows why my idea doesn't work?

Teacher. Yes. Look at

$$x^2 + 2x + y^2 + 2y + xy = 1.$$

Adding 1 to each of $x^2 + 2x$ and $y^2 + 2y$ leads to

$$(x + 1)^2 + (y + 1)^2 + xy = 3.$$

Now, does this mean $h = -1$ and $k = -1$? Not at all, since in the recipe for translating, one subtracts h from *each* occurrence of x, and k from each occurrence of y. *Completing the squares doesn't touch the xy-term.* And if we insist upon replacing x and y by $(x + 1)$ and $(y + 1)$ in xy, we then reintroduce x- and y-terms.

Student. I see. I guess there's no easy way out.

Scaling and Tilt

Teacher. Now let's assume we've applied (9.7) to put the conic's center at the origin. We once again observe that no selection of h and k ever affects A, B or C, so to find the angle of tilt we need consider only the equation $Ax^2 + Bxy + Cy^2 = 1$.

Interestingly, it turns out that when the conic is an *ellipse* in $\mathbb{R}^2$, adding linear terms $Dx + Ey$ not only shifts the ellipse, it also makes it larger.

> Suppose $Ax^2+Bxy+Cy^2 = 1$ defines a real ellipse. Adding any nonzero $Dx + Ey$ to $Ax^2 + Bxy + Cy^2$ always:
>
> - Shifts the ellipse;
> - Makes it larger;
> - Maintains the same shape;
> - Maintains the same tilt.

For any ellipse, the constant G in $Ax^2 + Bxy + Cy^2 = G$ acts in a way similar to a circle's radius squared: when it increases, the ellipse uniformly expands. In general, taking $G = 1$ may change the size of the conic, but that doesn't change its tilt, which is what we're interested in right now. That's why we can assume our conic is given by $Ax^2 + Bxy + Cy^2 = 1$.

Philosopher. But suppose G in (9.8) turns out to be zero? Then you can't divide by G to attain $G = 1$.

Teacher. Well, we're assuming that the conic is an ellipse or hyperbola—in other words, the conic isn't degenerate. That means G can't be zero; it's not too hard to show that $G = 0$ implies that the conic is in fact degenerate. One specific example is $x^2 + 2y - y^2 = 1$, which defines two crossing lines.

Now in the three coefficients A, B and C of $Ax^2 + Bxy + Cy^2 = 1$, there lurk three pieces of geometric information: shape, size and tilt, and our job is to extract this information. Here's our basic approach:

> *Any ellipse or hyperbola is the image of the unit circle $x^2 + y^2 = 1$ under some linear transformation defined by a nonsingular symmetric matrix S.*
> ***Find S in terms of A, B, and C.*** (9.9)

We've already begun translating our tilt/scale problem into the compact notation of linear algebra. The unit circle is our starting point, and it is time to translate its equation $x^2 + y^2 = 1$. Now (x, y) is a vector or row matrix, so we can write the circle equation as

$$(x, y) \cdot (x, y)^T = 1,$$

where T denotes transpose. By our action-reaction principle, to get the equation of the image of the circle under such a matrix S, replace every occurrence of (x, y) by $(x, y)S^{-1}$. Thus the linear-algebra equation of our ellipse is

$$(x, y)S^{-1} \cdot [(x, y)S^{-1}]^T = 1.$$

That isn't going to win any prizes for beauty, but we can neaten it up somewhat:

$$(x, y)S^{-1}(S^{-1})^T(x, y)^T = 1.$$

We can in fact do even better: since S symmetric, so is S^{-1}—that is, $(S^{-1})^T = S^{-1}$. That allows us to write our equation as

$$(x, y)S^{-2}(x, y)^T = 1. \tag{9.10}$$

This, of course, is just another way of writing $Ax^2 + Bxy + Cy^2 = 1$, which itself can be cast into the language of linear algebra:

$$(x, y)\begin{pmatrix} A & \frac{B}{2} \\ \frac{B}{2} & C \end{pmatrix}(x, y)^T = 1.$$

Due to its importance, let's make it official:

Denote this "coefficient matrix" S^{-2} by Λ.

Now Λ is not only symmetric, its entries are also real. Since the eigenvalues λ of a square matrix M are the roots of $\det(M - \lambda I) = 0$—a polynomial of degree n if M is $n \times n$—there are, by the fundamental theorem of algebra, n such zeros, counted with multiplicities. When M is a real, symmetric matrix S, we have this piece of linear algebra folklore:

> **A real nonsingular symmetric $n \times n$ matrix has n real nonzero eigenvalues counted with multiplicities, and n mutually orthogonal eigenvectors.** (9.11)

This leads to a simple geometrical statement:

> **A real nonsingular symmetric $n \times n$ matrix independently scales, without rotation, n mutually perpendicular 1-subspaces of $\mathbb{R}^n$.** (9.12)

Such matrices are therefore essentially just diagonal matrices (with no diagonal entries zero), except that the orthogonal directions of stretching may not necessarily lie along the *coordinate* directions.

Philosopher. This is fine news about Λ, but our concern is with S, not Λ. You said an ellipse or hyperbola is the image of a circle under S, and Λ didn't figure in this.

Teacher. Yes, but $\Lambda = S^{-2}$, and because Λ is geometrically so simple, going in reverse to get S is pretty easy. I'm proposing that starting from Λ we find S, in essence "solving for S in terms of Λ." Symbolically, $S = \Lambda^{-1/2}$, and because Λ is so uncomplicated in what it does geometrically, we can figure out what powers of Λ do, as well. For example, if Λ stretches along orthogonal 1-subspaces l_i by factors λ_i, then Λ^2 stretches along the same 1-subspaces l_i by the factors λ_i^2. Similarly, Λ^{-1} stretches along l_i by λ_i^{-1}, and $\Lambda^{1/2}$ stretches along l_i by $\lambda_i^{1/2}$. All this is not much different from multiplying diagonal matrices elementwise.

Philosopher. Aha! Those are interesting examples, but I'll bet the *real* enchilada is that $S = \Lambda^{-1/2}$ stretches along those orthogonal 1-subspaces l_i by $\lambda_i^{-1/2}$.

Teacher. Yes—that's the crucial insight.

Philosopher. So from any particular $Ax^2 + Bxy + Cy^2 = 1$ we get the coefficient matrix Λ, and from Λ, we figure out the directions and amounts of stretch. The stretch directions of S are the same, and its stretch factors are

$$s_1 = \frac{1}{\sqrt{\lambda_1}} \quad \text{and} \quad s_2 = \frac{1}{\sqrt{\lambda_2}}.$$

Since the ellipse is the image of the unit circle under S, we'll know what the ellipse looks like!

Student. That may be the game plan, but I feel as if I'm in a vacuum. For any of this to work, you need some way of getting those directions and stretch factors of Λ.

Teacher. Right, but we can continue to capitalize on the geometric nature of Λ. In any stretch direction (x, y), Λ just stretches by some factor λ, and one can express this, too, in the notation of linear algebra: $(x, y)\Lambda = (x, y)\lambda$. We can rewrite the right side of this so it has the same multiplicative structure as the left—that is, a (1×2)-matrix times a (2×2)-matrix. We do this by replacing the scalar λ by the (2×2)-matrix

$$\lambda I = \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix},$$

and we can then rearrange the equation to read

$$(x, y)(\Lambda - \lambda I) = (0, 0). \tag{9.13}$$

Now we can apply an analogy with ordinary real numbers, where the product $xc = 0$ has a nonzero solution only if the constant c is zero. In (9.13), "the constant c is zero" is taken in the sense that $\det(\Lambda - \lambda I) = 0$. The solutions (x, y) are the eigenvectors, and I like to call the lines on which they lie, "*eigenlines*." They're the lines along which we stretch. We thus have the ellipse's tilt.

We can also find the stretch factors λ_i of Λ, and therefore of $S = \Lambda^{-1/2}$ We've seen that for S, the stretches are

$$s_1 = \frac{1}{\sqrt{\lambda_1}} \quad \text{and} \quad s_2 = \frac{1}{\sqrt{\lambda_2}}$$

—the lengths of the semi-major and semi-minor axes of the ellipse.

Student. It seems that we've pretty much solved the problem!

Philosopher. I find it very interesting that earlier, you mentioned the quadratic formula as a good example of "going in reverse" from the algebra of a quadratic equation to the geometry of sketching its roots. It seems that at the heart of reversing things at this more general level, we once again encounter a quadratic equation: $\det(\Lambda - \lambda I) = 0$ is quadratic in λ, and one can use the quadratic formula to solve it explicitly.

Teacher. Quadratic equations and the quadratic formula amount to a theory of "conics in 1-space." For example, we can easily restrict to one dimension by setting either x or y equal to a constant in (9.3) [p. 147], and that turns (9.3) into an ordinary one-variable quadratic equation.

Philosopher. That puts a nice perspective on things.

Student. After all this talk, an example would definitely help put things in better focus for me.

Teacher. It is indeed time.

Let's make up a tilted ellipse centered at the origin, find its equation, and then apply what we've learned in order to recapture the original ellipse. That would have a nice self-check built into it. In fact, since we've already worked out the equation of a 2×1 ellipse rotated counterclockwise 30°, let's use it:

$$\frac{7}{16}x^2 - \frac{3\sqrt{3}}{8}xy + \frac{13}{16}y^2 = 1.$$

This gives

$$\Lambda = \begin{pmatrix} \frac{7}{16} & \frac{-3\sqrt{3}}{16} \\ \frac{-3\sqrt{3}}{16} & \frac{13}{16} \end{pmatrix}.$$

The eigenvalues and corresponding eigenvectors turn out to be

$$\lambda_1 = \frac{1}{4}, \qquad v_1 = (\sqrt{3}, 1);$$

$$\lambda_2 = 1, \qquad v_2 = (-1, \sqrt{3}).$$

Of course v_1 lies along the 30°-line and v_2 lies along the 120°-line. Since $\Lambda = S^{-2}$, the eigenlines of S are those of Λ, and the stretch factors of S are

$$s_1 = \frac{1}{\sqrt{\lambda_1}} = 2 \quad \text{and} \quad s_2 = \frac{1}{\sqrt{\lambda_2}} = 1.$$

This gives us exactly the ellipse we started with.

Student. Beautiful!

Hyperbola

Student. I've been thinking about our basic strategy in (9.9) [p. 151]:

> *Any ellipse or hyperbola is the image of the unit circle $x^2 + y^2 = 1$ under some nonsingular symmetric linear transformation. Find that transformation from A, B, and C.*

It doesn't seem that this approach works for hyperbolas. I've tried and tried, but every linear transformation I use—even nonsymmetric ones like rotations or shears—produces an ellipse, never a hyperbola.

Philosopher. At the beginning of our journey, I would have been very upset at what you're saying. But so many times we've hung tight, insisting on unity, and we have so far prevailed. Certainly when it comes to standard forms, every hyperbola can be looked at as an ellipse. For example

$$\left(\frac{x}{a}\right)^2 - \left(\frac{y}{b}\right)^2 = 1 \quad \text{is} \quad \left(\frac{x}{a}\right)^2 + \left(\frac{y}{ib}\right)^2 = 1.$$

And it has been pretty well branded onto my brain that those denominators are stretching factors! So here goes: "In Mathematics we trust"—I'm going to say that the hyperbola

$$\left(\frac{x}{a}\right)^2 - \left(\frac{y}{b}\right)^2 = 1$$

is the unit circle $x^2 + y^2 = 1$ stretched horizontally by a and vertically by ib.

Student. But are we allowed to stretch by ib?

Teacher. The only restriction we put on scaling is that it's nonzero. Since b isn't zero, ib isn't either.

Student. OK, but I sure don't see this geometrically!

Philosopher. I don't either—not right now. But my feeling is there's a picture to be had.

Teacher. I don't think it will be so hard to draw a picture of what's going on. Since we're dealing with standard forms, the eigenlines are just the x- and y-axes, and we can opt to look at these as complex axes. Then every vector on the x-axis is scalar multiplied by a, and every vector on the y-axis, by ib. This last is what's interesting, because "complex scaling" by ib appears visually as ordinary uniform scaling of the complex y-plane by b, followed by a 90° rotation about its origin. That rotation of course comes from the i. So I propose this: take the unit circle and stretch by *real* amounts a and b, and then do the 90° rotation in the complex y-axis to account for the i in ib. Since we're limited to three real dimensions, let's look in (x_1, y_1, iy_2)-space, where we can see the real stretches by a and b, as well as the i-induced rotation by 90° in (y_1, iy_2)-space. Here are pictures showing the two steps:

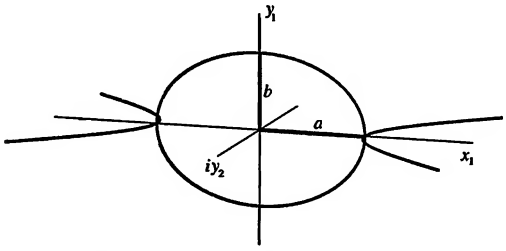


FIGURE 9.1. This is the unit circle $x^2+y^2 = 1$ stretched by a in the x-direction and by real b in the y-direction. Now multiply b by i . . .

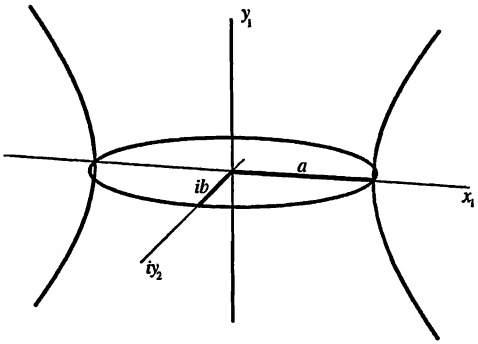


FIGURE 9.2. ...and notice how this rotates the ellipse/hyperbola in Figure 9.1 by 90° about the x_1-axis. What one sees in the ordinary (x_1, y_1)-viewing screen has changed from an ellipse to a hyperbola.

Student. I'm impressed! But I still must be missing something. Why couldn't I ever get a hyperbola? Come to think of it, what linear transformation *are* you using to make our basic philosophy continue to work?

Teacher. You're asking what nonsingular symmetric matrix S am I applying to the unit circle to make it a hyperbola?

Student. Yes...

Teacher. Take a specific hyperbola, like $x^2 - y^2 = 1$. Its real symmetric coefficient matrix is

$$\Lambda = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Writing $\Lambda = S^{-2}$ as we did earlier means that $S = \Lambda^{-1/2}$. This can be computed elementwise since Λ is diagonal; one such matrix is

$$S = \begin{pmatrix} 1 & 0 \\ 0 & i \end{pmatrix}.$$

Student. That *is* nonsingular and symmetric... but...

Philosopher. ...but it isn't real! No wonder you couldn't find a linear transformation—you were looking for a *real* one.

Teacher. But notice that the wording of our basic philosophy in (9.9) [p. 151] allows us to apply that philosophy as is, since it doesn't say the symmetric matrix has to be real.

Philosopher. It appears the big tip-off is noticing whether one of the eigenvalues of Λ is negative. That's when we get a hyperbola. We know Λ is real and symmetric, so its eigenvalues are real, but that doesn't rule out negative eigenvalues! When some λ_i is negative,

we get an imaginary stretch factor in S, since the eigenvalues of S are

$$s_i = \frac{1}{\sqrt{\lambda_i}}.$$

> In the real setting, a conic in standard form intersects a coordinate axis exactly when the corresponding coefficient in the equation is positive; this works in any number of variables. Thus, for example, $3x^2 - 4y^2 + z^2 = 1$ intersects the x- and z-axes, but not the y-axis, and we have a hyperboloid of one sheet. $-x^2 - y^2 + z^2 = 1$ intersects only the z-axis, and is a hyperboloid of two sheets. $2x^2 + y^2 + 3z^2 = 1$ intersects all three axes and is an ellipsoid. $-x^2 - 2y^2 - z^2 = 1$ intersects no axes and defines the empty set.

Teacher. Yes! An imaginary stretch factor does to an arbitrary complex eigenline just what it does to the complex y-axis—uniformly real-stretches it, and then rotates it 90° about the origin. Figures 9.1 and 9.2 are actually quite typical of what's going on in the more general case.

An Off-the-Wall Question

Philosopher. Another question has entered my head. Maybe it's somewhat crazy, but it would be nice if we could answer it. If a and b are real, then what about an equation like

$$\left(\frac{x}{ia}\right)^2 + \left(\frac{y}{ib}\right)^2 = 1?$$

Here *both* stretching factors are imaginary. This equation can be written as

$$\left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 = -1,$$

showing that it doesn't have any real solutions. Does this mean that rotating *both* the complex x-axis and complex y-axis by 90° moves the complex conic

$$\left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 = 1$$

in such a way that it entirely misses the ordinary real viewing screen?

Teacher. That sounds plausible. Let's check it by starting with Figure 9.2, where we've already performed one rotation, which represents going from

$$\left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 = 1$$

to

$$\left(\frac{x}{a}\right)^2 + \left(\frac{y}{ib}\right)^2 = 1.$$

For the second rotation, look at

$$\left(\frac{x}{a}\right)^2 - \left(\frac{y}{b}\right)^2 = 1$$

in (x_1, ix_2, y_1)-space (the 3-space defined by $y_2 = 0$) and rotate the points of the complex x-axis (that is, the (x_1, ix_2)-plane) counterclockwise through an angle of 90° to account for the i in ai. To see the part of

$$\left(\frac{x}{a}\right)^2 - \left(\frac{y}{b}\right)^2 = 1$$

in (x_1, ix_2, y_1)-space, we can use the sequence of equations leading to Figure 6.1 [p. 92]. Here, however, we're taking $y_2 = 0$ instead of $x_1 = 0$. From this, we deduce that either $x_1 = 0$ or $x_2 = 0$, and this leads to the two equations

$$-\left(\frac{x_2}{a}\right)^2 - \left(\frac{y_1}{b}\right)^2 = 1 \text{ (when } x_1 = 0), \tag{9.14}$$

$$\left(\frac{x_1}{a}\right)^2 - \left(\frac{y_1}{b}\right)^2 = 1 \text{ (when } x_2 = 0), \tag{9.15}$$

and in (x_1, ix_2, y_1)-space these define two loci: (9.14) defines the empty set, and (9.15), a hyperbola. Here's the sketch:

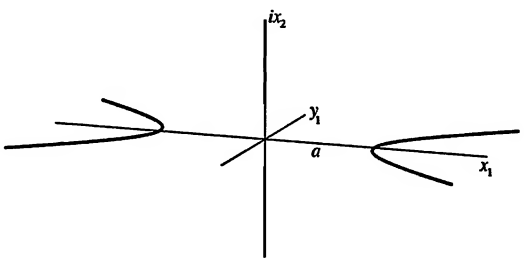


FIGURE 9.3. In (x_1, ix_2, y_1)-space, this is all we see of the conic

$$\left(\frac{x}{a}\right)^2 - \left(\frac{y}{b}\right)^2 = 1.$$

The hyperbola lies in the real (x_1, y_1)-plane. Rotating points of the complex x-axis counterclockwise through an angle of 90° shows up in this picture as rotation of the hyperbola about the y_1-axis, leaving...

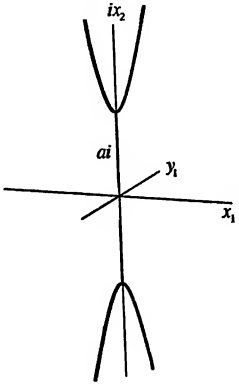


FIGURE 9.4. ...nothing visible in the (x_1, y_1)-plane.

Your conjecture is correct!

An Example

Student. I would like to see an example of a tilted hyperbola, and verify that our method still works.

Teacher. Yes, that's a good idea. Why not try a 2-by-1 hyperbola rotated counterclockwise by 30°, as a sort of companion to the ellipse example? Actually, we can use formula (9.1) [p. 145] as is, since it was derived merely by rotating the standard form of an ellipse. When one of a, b is imaginary we happen to call the conic a "hyperbola," but that in no way changes (9.1)—just replace a by ai or b by bi as the case may require. Since a and b appear in (9.1) only as their squares, all coefficients in the formula are real.

Student. To get a hyperbola, we can, for example, substitute $a = 2$, $b = i$ and $\theta = 30°$ into (9.1), right?

Teacher. Sure. That gives this companion to (9.2):

$$\frac{-1}{16}x^2 + \frac{5\sqrt{3}}{8}xy - \frac{11}{16}y^2 = 1. \tag{9.16}$$

To go in reverse—that is, to figure out what θ leads to (9.16), we start by writing

$$\Lambda = \begin{pmatrix} \frac{-1}{16} & \frac{5\sqrt{3}}{16} \\ \frac{5\sqrt{3}}{16} & \frac{-11}{16} \end{pmatrix}.$$

The eigenvalues and corresponding eigenvectors are

$$\lambda_1 = \frac{1}{4}, \qquad v_1 = (\sqrt{3}, 1);$$

$$\lambda_2 = -1, \qquad v_2 = (-1, \sqrt{3}).$$

As we'd expect, the eigenlines are the same as in the corresponding ellipse example. The symmetric matrix S that transforms the unit circle therefore has stretch factors

$$s_1 = \frac{1}{\sqrt{\lambda_1}} = 2 \quad \text{and} \quad s_2 = \frac{1}{\sqrt{\lambda_2}} = \frac{1}{i} = -i.$$

Philosopher. Is that minus sign a problem?

Teacher. No, since replacing x and y by their negatives does nothing to

$$Ax^2 + Bxy + Cy^2 = 1.$$

Phrased geometrically, conics centered at the origin are symmetric about the origin, so any stretch factor

$$s_i = \pm\frac{1}{\sqrt{\lambda_i}}$$

does the same thing to the unit circle as

$$-s_i = \mp\frac{1}{\sqrt{\lambda_i}}.$$

The Discriminant Discriminates

Student. Just looking at equation (9.2) [p. 146] or equation (9.16), I get no clue at all as to whether we've got an ellipse or a hyperbola. But the standard form reveals the answer immediately. Is there some way, short of solving the entire problem, to see from (9.2) or (9.16) what kind of conic we have?

Teacher. Yes, there is. In the standard forms it's easy to see that a real, nonempty conic is an ellipse if and only if both eigenvalues λ_i of

$$\Lambda = \begin{pmatrix} \frac{1}{a^2} & 0 \\ 0 & \frac{1}{b^2} \end{pmatrix}$$

are positive (in which case $\lambda_1\lambda_2 > 0$), and a hyperbola if and only if one is negative and the other positive (that is, if $\lambda_1\lambda_2 < 0$). Now $\lambda_1\lambda_2$ is the product of the roots of this quadratic equation in λ:

$$\det(\Lambda - \lambda I) = 0; \tag{9.17}$$

$\lambda_1\lambda_2$ is the constant term of the left side. This term is also the result of setting the variable λ equal to 0 in (9.17), so we can write

$$\lambda_1\lambda_2 = \det(\Lambda) = \det\begin{pmatrix} A & \frac{B}{2} \\ \frac{B}{2} & C \end{pmatrix} = AC - \frac{B^2}{4}. \tag{9.18}$$

Philosopher. $AC - \frac{B^2}{4}$ has a familiar look to it.

Student. It looks something like $B^2 - 4AC$.

Teacher. Yes! The discriminant $B^2 - 4AC$ is $-4(AC - \frac{B^2}{4})$. So in the ellipse case, $\lambda_1\lambda_2$ and $AC - \frac{B^2}{4}$ are both positive, meaning that the discriminant is negative. In the hyperbola case, the discriminant is positive. So the discriminant of $Ax^2 + Bxy + Cy^2$ easily decides for us.

How the Discriminant Discriminates

Suppose that $Ax^2 + Bxy + Cy^2 = 1$ defines a nonempty conic in $\mathbb{R}^2$. This curve is symmetric about the origin and is

- An ellipse if $B^2 - 4AC < 0$;
- Two parallel lines if $B^2 - 4AC = 0$;
- A hyperbola if $B^2 - 4AC > 0$.

Area

Philosopher. The area of the ellipse

$$\left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 = 1$$

is πab. Now the ellipse $Ax^2 + Bxy + Cy^2 = 1$ is completely determined by A, B, and C, so it seems to me there ought be an area formula directly in terms of these coefficients.

Teacher. That certainly sounds plausible.

Philosopher. But to *find* it!

Teacher. I think it would help to interpret the πab formula geometrically—it's just π times the product of the semi-axis lengths. That formula could be written on a cardboard ellipse, and the ellipse could be rotated as much as you wish. We know that the semi-axes of $Ax^2 + Bxy + Cy^2 = 1$ are

$$\frac{1}{\sqrt{\lambda_1}} \quad \text{and} \quad \frac{1}{\sqrt{\lambda_2}},$$

so the ellipse has area

$$\frac{\pi}{\sqrt{\lambda_1\lambda_2}}.$$

Philosopher. OK, but that's not in terms of A, B and C.

Teacher. But (9.18) *exactly* expresses $\lambda_1\lambda_2$ in terms of A, B and C! The area is therefore

$$\frac{\pi}{\sqrt{AC - \frac{B^2}{4}}}.$$

We can rewrite this a bit to bring out the nice discriminant aspect:

$$\text{Area} = \frac{2\pi i}{\sqrt{B^2 - 4AC}}. \tag{9.19}$$

Philosopher. This is nice. I notice that for an ellipse, B^2-4AC is negative, so $\sqrt{B^2 - 4AC}$ is imaginary. That cancels out the i in the numerator, giving a real answer.

Student. Yes, but suppose you've got a hyperbola. Its discriminant is positive! Then what? You're getting nonsense—an imaginary answer, because the i doesn't cancel. Just looking at a hyperbola, it appears that its "area" is infinite. Our formula doesn't seem to have anything to do with a hyperbola.

Philosopher. But remember how fortunate we've been when we preserve the algebraic form? We see that (9.19) works for an ellipse. No matter what answer it gives us, I think it's trying to tell us something. Maybe it will lead us to some "deeper truth" that we've been missing. We should listen!

Student. OK, suppose you're right. How are you going to preserve the form of πab?

Philosopher. One of our great insights has been that we can look at a hyperbola in standard form as an ellipse in standard form. So let's write

$$\left(\frac{x}{a}\right)^2 - \left(\frac{y}{b}\right)^2 = 1 \quad \text{as} \quad \left(\frac{x}{a}\right)^2 + \left(\frac{y}{ib}\right)^2 = 1.$$

Student. But then the standard-form area formula gives us $i\pi ab$.

Philosopher. That's imaginary! Well, at least things are consistent.

Teacher. We'd better look at a picture, to see what's going on.

We see the ordinary hyperbola in the (x_1, y_1)-plane, but in the (x_1, iy_2)-plane, we see an ellipse with semi-axes a and ib. The units of measurement along the coordinate axes of this Minkowski plane are 1 and i, and that makes the unit of area $1 \cdot i = i$. It's not surprising that we get imaginary areas. And π times the product of the semi-axes of this ellipse is indeed $i\pi ab$. It appears that preserving the algebraic form works very nicely. It fact, it gives our formula (9.19) more power than we realized—it leads us to the idea of area in the Minkowski plane. [See Chapter 6, exercise 4 on p. 99.]

Student. I would really like to see some examples of how the formula works in nonstandard forms.

Teacher. You keep us honest! Let's apply the formula to each of the companion examples we've already studied.

First, let's look at the tilted ellipse

$$\frac{7}{16}x^2 - \frac{3\sqrt{3}}{8}xy + \frac{13}{16}y^2 = 1$$

in (9.2). We got this equation by tilting a 2×1-ellipse 30°

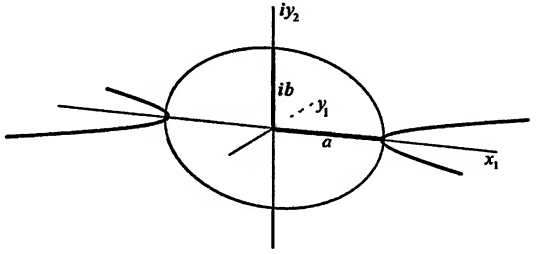


FIGURE 9.5. Writing the hyperbola

$$\left(\frac{x}{a}\right)^2 - \left(\frac{y}{b}\right)^2 = 1$$

in the form $Ax^2 + Bxy + Cy^2 = 1$ means

$$A = \frac{1}{a^2}, B = 0 \quad \text{and} \quad C = -\frac{1}{b^2}.$$

The area formula

$$\frac{2\pi i}{\sqrt{B^2 - 4AC}}$$

then reduces to πabi. Just as we do, this formula regards the hyperbola as the ellipse

$$\left(\frac{x}{a}\right)^2 + \left(\frac{y}{ib}\right)^2 = 1,$$

an ellipse with one imaginary principal axis.

Student. Tilting shouldn't change the area. Since it's a 2-by-1 ellipse, the area ought to be $\pi \cdot 2 \cdot 1 = 2\pi$.

Teacher. Right, so we've got an acid test. Now

$$A = \frac{7}{16}; \quad B = -\frac{3\sqrt{3}}{8}; \quad C = \frac{13}{16}.$$

The discriminant is therefore

$$B^2 - 4AC = \frac{27}{64} - 4 \cdot \frac{7}{16} \cdot \frac{13}{16} = \frac{27 - 91}{64} = \frac{-64}{64} = -1.$$

This leads at once to $Area = 2\pi$.

Student. Impressive! Now try the hyperbola example.

Teacher. Equation (9.16) is

$$\frac{-1}{16}x^2 + \frac{5\sqrt{3}}{8}xy - \frac{11}{16}y^2 = 1,$$

and its locus is a tilted 2-by-1 hyperbola, but looked at as an ellipse, its semi-axes are 2 and i. That means its area should be $2\pi i$—another acid test. We have:

$$A = -\frac{1}{16}; \quad B = \frac{5\sqrt{3}}{8}; \quad C = -\frac{11}{16}.$$

This has discriminant

$$B^2 - 4AC = \frac{75}{64} - 4 \cdot \frac{-1}{16} \cdot \frac{-11}{16} = \frac{75 - 11}{64} = \frac{64}{64} = 1,$$

which indeed produces an area of $2\pi i$.

Student. I'm getting more and more convinced about this "preserving the form" idea. It really seems to work.

Teacher. That one simple principle has done so much. It has predicted hidden foci and directrices, too!

> "Newly-created mathematics is like a fresh snowfall—you see inches and inches of white stuff, and it's intimidating. Then, with time, it all melts down to maybe less than an inch of water, and you wonder what all the fuss was about!"
>
> Harish-Chandra

Chapter Summary

Starting with the unit circle $x^2 + y^2 = 1$, we can obtain any ellipse geometrically by first stretching in the x- and/or y-direction, then rotating about the origin, and finally parallel-translating. These transformations have simple ("action-reaction") algebraic counterparts, and that means this process yields the ellipse's equation. We can obtain a hyperbola by stretching the circle in one direction by an imaginary quantity. The remainder of the chapter is mostly devoted to reversing this geometry-to-algebra process. That is, we tease out the geometric information lurking in a quadratic form. Like peeling layers off an onion, we first show how to "unshift" so the conic is centered at the origin. We then "unrotate" and "unstretch" to arrive at the circle $x^2 + y^2 = 1$. After the "unshift" step—centering at the origin—the problem reduces to extracting the remaining three pieces of geometric information—shape, size and tilt—from the three pieces of algebraic data A, B, C in $Ax^2 + Bxy + Cy^2 = 1$. The strategy for obtaining that geometric information is encapsulated in (9.9) on p. 151, and we use standard "eigen-arguments" to carry it out.

Exercises

1. Use the "action-reaction" principle to rotate the ellipse

$$\left(\frac{x}{3}\right)^2 + \left(\frac{y}{5}\right)^2 = 1$$

clockwise through 60° Then use eigen-arguments to find its semi-axes and angle of tilt. Show that the formula in (9.19) on p. 162 gives the same area before and after rotating.

2. (a) The hyperbola in (9.16) on p. 159 came from the ellipse in (9.2) on p. 146 by replacing $b = 1$ by $b = i$. The "other" hyperbola is obtained by instead replacing $a = 2, b = 1$ in (9.2) by: $a = 2i, b = 1$. You are surprised to find that the Λ for this new hyperbola is the negative of the Λ for the hyperbola in (9.16). Is this a fluke, or will this always happen when you switch the position of i?
   (b) Suppose you change a to ia *and* B to ib. What's the relation of the new Λ to the original one?
   (c) Supply a geometric rationale for your observations.
3. Apply the area formula in (9.19) to the conic in (9.1) on p. 145, showing that the big expression finally reduces to πab.
4. Verify the chapter's assertion about decimal-expansion repetition in the case of $\frac{1}{M}$, for $M = 2, \ldots 12$.
5. Prove that in the real setting, shifting an ellipse $Ax^2 + Bxy + Cy^2 = 1$ to $Ax^2 + Bxy + Cy^2 + Dx + Ey = 1 (D \cdot E \neq 0)$ always makes the ellipse larger.
6. If h and k are as in (9.7) (p. 150), and if in Exercise 5 we instead assume that $B^2 - 4AC > 0$, then prove:
   (a) if (h, k) lies in the first or third quadrant of the (x, y)-plane, then the hyperbola $Ax^2 + Bxy + Cy^2 + Dx + Ey = 1$, in addition to being shifted, is also magnified (we zoom in);
   (b) if (h, k) lies in the second or fourth quadrant of the (x, y)-plane, then the hyperbola $Ax^2 + Bxy + Cy^2 + Dx + Ey = 1$ is not only shifted, but our view of it is zoomed *out*;
   (c) if (h, k) lies on an *axis* of the (x, y)-plane, then we have pure shift, with no zooming.
7. In Exercise 5, "always makes the ellipse larger" sounds suspiciously arbitrary, and would surely upset Philosopher. Is there some way to restate things and round out the story to bring about a more symmetric statement?
8. For each of the following curves in $\mathbb{R}^2$, draw a typical picture exemplifying the condition and find a general equation corresponding to each picture. Is the discriminant zero in each of these cases?
   (a) Two distinct, crossing lines;
   (b) two parallel lines;
   (c) one double line (If $f(x, y) = 0$ is the equation of a curve, then $(f(x, y))^2 = 0$ is the equation of the double curve.);
   (d) a parabola;
   (e) a point.
9. **"Three Gems."** Since any conic centered at the origin has equation

$$Ax^2 + Bxy + Cy^2 = 1,$$

there ought to be formulas in terms of A, B, C directly giving its "vital stats" such as size, shape, tilt, and area. We've already met one of three "gemlike" equations,

$$\text{Area} = \frac{2\pi i}{\sqrt{B^2 - 4AC}}.$$

In (a) and (e) next, we meet two more.

(a) We've seen that the eigenvalues λ_i determine size and shape through the stretching factors

$$\frac{1}{\sqrt{\lambda_i}}.$$

Since

$$x = \frac{1}{\sqrt{\lambda}} \quad \text{implies that} \quad \lambda = \frac{1}{x^2},$$

instead of subtracting λ from the diagonal entries of the symmetric matrix

$$\begin{pmatrix} A & \frac{B}{2} \\ \frac{B}{2} & C \end{pmatrix},$$

why not directly subtract $\frac{1}{x^2}$ *to get a polynomial whose zeros are the stretch factors?* Do this, thereby showing that the stretch factors are the roots of this quadratic in x^2:

$$(B^2 - 4AC)x^4 + 4(A + C)x^2 - 4 = 0.$$

(b) Verify that the formula in (a) gives the correct answer for the hyperbola $x^2 - y^2 = 1$.

(c) Apply the formula in (a) to $x^2 + 2xy + y^2 = 1$. Notice that this formula does not give the "∞" stretch factor. But geometrically, such a stretch factor means something. *What*?

(d) If $B^2 - 4AC \neq 0$, you can divide through by it to make the polynomial in part (a) monic. Then its constant term is the product of its zeros. Use this idea to obtain another derivation of the area formula

$$\text{Area} = \frac{2\pi i}{\sqrt{B^2 - 4AC}}.$$

(e) A conic's tilt is determined by eigenvectors, but actually only an eigenvector's *slope* is ever used, and that is only a scalar. The question therefore arises: might there be a polynomial whose zeros directly give the eigenline slopes m? There is. The slopes are the roots of

$$Bm^2 + 2(A - C)m - B = 0.$$

Prove this. Notice that when $B = 0$, the equation yields the solution $m = 0$, but not $m = \infty$. [One possible approach to proving the formula: solve for the m_i; then

two eigenvectors are $(1, m_i)$. Show that multiplying either of these eigenvectors by the matrix

$$\begin{pmatrix} A & \frac{B}{2} \\ \frac{B}{2} & C \end{pmatrix}$$

does in fact merely scalar multiply each $(1, m_i)$. In terms of A, B, C, what are these two scalar multipliers (stretch factors)?]

10. Since area is unaffected by rotation and the area of the conic $Ax^2 + Bxy + Cy^2 = 1$ is

$$\frac{2\pi i}{\sqrt{B^2 - 4AC}},$$

we easily see that the discriminant of $Ax^2 + Bxy + Cy^2$ is invariant under rotation. Show that likewise, $A + C$ is invariant under rotation. Is there a formula in this chapter that makes this obvious? Are there two such formulas?

11. Use the "action-reaction principle" to find the equation of the parabola $y = x^2$ after it's been rotated clockwise 30° about the origin.

12. We've seen in this chapter that if $\lambda_1 > 0$ and $\lambda_2 > 0$, then the real conic is an ellipse, while if $\lambda_1\lambda_2 < 0$, it's a hyperbola. We've also seen from the box on p. 157 that a conic in standard form intersects a coordinate axis exactly when the corresponding coefficient in the equation is positive. How are these two observations related? Geometrically, what does the sign of each λ_i tell us?

13. Teacher states (on p. 151) that if G in (9.8) turns out to be zero, then the conic must be degenerate. Prove this.

14. From what you've learned from Exercise 9, and from what you know about the discriminant, about roots of general polynomial equations, and about stretching factors, explain why it makes sense that the constant term in

$$(B^2 - 4AC)x^4 + 4(A + C)x^2 - 4 = 0$$

is negative.

15. If $B \neq 0$ in $Ax^2 + Bxy + Cy^2 = 1$, then we can write the equation

$$Bm^2 + 2(A - C)m - B = 0$$

in Exercise 9(e) as

$$m^2 + 2\frac{(A - C)}{B}m - 1 = 0.$$

Why is it sensible that the constant term is -1?

16. When $B \neq 0$ in $Ax^2 + Bxy + Cy^2 = 1$, then

$$Bm^2 + 2(A - C)m - B = 0 \quad \text{and} \quad (B^2 - 4AC)x^4 + 4(A + C)x^2 - 4 = 0$$

give the slopes and stretch factors of the conic $Ax^2 + Bxy + Cy^2 = 1$, and all values are $\neq 0, \neq \infty$. (Why?) However, these formulas don't completely determine the conic itself, because they don't tell us which stretch factor corresponds to which slope. The following formulas fill in the "missing link."

The stretch factor corresponding to slope m is

$$\sqrt{\frac{1+m^2}{A+Bm+Cm^2}}.$$

The stretch factor corresponding to slope $-\frac{1}{m}$ is

$$\sqrt{\frac{1+m^2}{Am^2-Bm+C}}.$$

Prove these formulas. Do they make sense for the conic $x^2 + 2xy + y^2 = 1$? For the conic $x^2 - 2xy + y^2 = 1$? Why?

17. If in $\mathbb{R}^2$ you subject the locus of $f(x, y) = 0$ first to a nonsingular linear transformation A, and then to a nonsingular linear transformation B, prove that in our "action-reaction principle," the change-of-variable is $(x, y) \to (x, y)B^{-1}A^{-1}$. Give a counterexample showing that $(x, y) \to (x, y)A^{-1}B^{-1}$ can give an incorrect result.

10

A Beautiful Mathematical Universe

in which removing further exceptions leads to an expanded world of great beauty and utility

A "Conservation Principle"

Philosopher. The image of tossing a wire ellipse onto the plane got me thinking. If one tosses *two* ellipses onto the plane, then all sorts of exceptions arise. Most of the time, the two ellipses fall far from each other and don't intersect. But in Figure 10.1, we see two ellipses intersecting in four points:

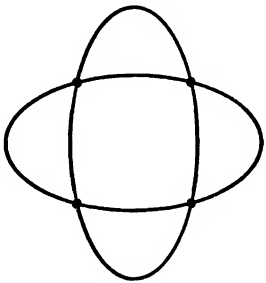


FIGURE 10.1. The two ellipses

$$\frac{x^2}{1^2}+\frac{y^2}{2^2}=1 \quad \text{and} \quad \frac{x^2}{2^2}+\frac{y^2}{1^2}=1$$

intersect in four points. It turns out that in an appropriate sense, any two distinct conics *always* intersect in four points.

Translating either of them shows they could also intersect in *two* points, or in no points at all.

Student. As you go from four to two points, don't they intersect in three? And as you go from two to zero, don't they intersect in one point?

Philosopher. Yes, but in each of those cases the ellipses are tangent, and I'm counting with multiplicity 2 a point of tangential intersection, in the same way that we do when the parabola $y = x^2$ tangentially intersects the x-axis. But zero, two or four points of intersection still represents a splintering into separate cases. It seems like mathematics versus Mathematics again—many islands instead of one unified mountain range!

Teacher. Let me pose a leading question. What is it that makes the fundamental theorem of algebra so powerful and beautiful? This theorem says that a polynomial of positive degree n *always* has n zeros, counted with multiplicities—no exceptions. And if we agree that a zero of multiplicity r intersects the x-axis in an r-fold point, then this theorem geometrically says that the x-axis intersects the graph in n points, counted with multiplicities. Now that, to me, seems powerful and beautiful, and maybe something similar is going on with ellipses—any two different ellipses should analogously intersect in a constant number of points.

Student. And you're thinking that number is *four*?

Philosopher. That's what *I* would guess. Certainly in the case when each conic is degenerate—say, each one consisting of two separate lines—then two lines intersect two other lines in a total of four points.

Student. I see that. Even if a line in one conic is parallel to a line the other conic, the two lines still intersect at infinity.

Philosopher. Right. I also believe that if one of the conics is nondegenerate, we *still* get four points.

Teacher. What's your reasoning?

Philosopher. Roughly this: a single line intersects a nondegenerate conic in two points, because relative to appropriate coordinates, any line can be parametrized by $x = a_1t + b_1$, $y = a_2t + b_2$, where a_1 and a_2 are nonzero. Then substituting $a_1t + b_1$ and $a_2t + b_2$ for x and y in the conic's equation almost always gives an equation of degree 2 in t. This has two solutions, counted with multiplicity, and these two t's give two points (x, y).

Teacher. Did I hear you say "almost always"? Might there be an exception?

Philosopher. I think one could remove that by appropriately changing coordinates.

Student. So a degenerate conic consisting of *two* lines intersects the other conic in $2 + 2 = 4$ points?

Philosopher. That's the idea. And if you then look at *two* nondegenerate conics, I think that if you stretch one of them enough, it'll still intersect the other in four points. That's because the stretched-out conic looks a lot like two parallel lines, and therefore they intersect in four points. Of course all this is "intuiting," but I'm trying to amass evidence for what I hope is true.

Teacher. It seems to me that if you start with two ellipses intersecting in four points, as in Figure 10.1, and begin to translate one of them, it's right after they're tangent that we lose two points.

Philosopher. Yes. As I said before, the situation seems similar to the parabola $y = x^2 - 1$, which intersects the x-axis in two points. In Figure 5.3 [p. 71], as we push the parabola upward, the intersection points approach each other, and do *not* disappear when the parabola rises above the x-axis. The coalesced points instead just split apart and separate in a different direction, along the imaginary part of the complex x-axis. This gives me a pious hope.

Student. Which is...

Philosopher. Well, I'm hoping this phenomenon occurs more generally. In Figure 10.1, for example, slowly push the tall ellipse upward. The bottom part of the other ellipse is somewhat like the x-axis in the parabola example. Here's a picture:

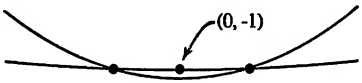


FIGURE 10.2. This is a magnified version of part of Figure 10.1 after the tall ellipse has been pushed up almost to the point of tangency.

It would be very nice if, right after the points coalesce, *they similarly separate along some "imaginary" axis.*

In Figures 5.2 and 5.3 on pp. 70 and 71, when we translate the tangent line at $P = (0, 0)$, we see two real intersection points coalesce and then "separate at right angles" along imaginary directions. This phenomenon holds somewhat more generally: one can replace the parabola by any curve C having, in some rectangular coordinate system (x, y), an even power series representation at P of the form $y = a_2x^2 + a_4x^4 + \cdots$, where $a_2 \neq 0$.

- Familiar examples: Translate a secant line of a real circle. When it becomes tangent at a point P, the real intersection points have just coalesced, and upon continued translation, they separate along imaginary directions in the moving complex secant line. The same thing happens at any maximum or minimum point of the (real) graphs of $y = \sin x$ or $y = \cos x$ as we translate a horizontal tangent line. Another example is the Gauss bell-shaped curve at its maximum.

  In fact, the tangent line itself can be replaced by any curve having a power-series representation as above, just as long as the two series are not identical.

- Examples: Take two circles tangent at P, and translate one of them along the common normal at P The double intersection point will separate into two distinct points moving in either the real or the imaginary direction. The same is true for two ellipses if they're tangent at some point of a common principal axis.

Teacher. So you're guessing that perhaps we just don't see them, just as we don't see the points of the pushed-up parabola on the x-axis, but that they're actually there?

Philosopher. That's my guess.

Teacher. I think we need to look at a specific example to see if what you're saying makes sense.

Student. An example would help to make this concrete.

Teacher. Let's continue looking at Figure 10.1. The tall ellipse there has equation

$$\frac{x^2}{1^2}+\frac{y^2}{2^2}=1,$$

and the other ellipse there has equation

$$\frac{x^2}{2^2}+\frac{y^2}{1^2}=1.$$

Substituting

$$x^2=1-\frac{y^2}{2^2}$$

from the first equation into the second and simplifying gives $5y^2=4$, so

$$y=\pm\frac{2}{\sqrt{5}}.$$

Substituting this into the equation of either ellipse gives

$$x=\pm\frac{2}{\sqrt{5}}.$$

The four intersection points are therefore

$$\left(\pm\frac{2}{\sqrt{5}},\pm\frac{2}{\sqrt{5}}\right),$$

where the two $\pm$'s are to be chosen independently. There are just two x-values and two y-values, and that is corroborated by the symmetric arrangement of the two ellipses.

Philosopher. So far, so good. Now suppose we push the taller ellipse upward.

Teacher. The equation of the tall ellipse, when translated upward by k, is

$$\frac{x^2}{1^2}+\frac{(y-k)^2}{2^2}=1.$$

If we eliminate y in this equation and in that of the stationary ellipse, we're led to this relation between x^2 and k:

$$k=\pm\sqrt{1-\frac{x^2}{4}}\pm 2\sqrt{1-x^2}, \tag{10.1}$$

where once again, the two ±'s are chosen independently. The x's associated with a given k are the x-coordinates of the intersection points of the fixed and translated ellipses. Notice that setting $x = 0$ gives four possible values of k: $-3, -1, 1, 3$. These are the four vertical translates where the ellipses become tangent. Here's the graph of k versus x^2 for

$$k = -\sqrt{1 - \frac{x^2}{4}} + 2\sqrt{1 - x^2},$$

describing the behavior of intersection points around the point $(0, -1)$ in Figure 10.2:

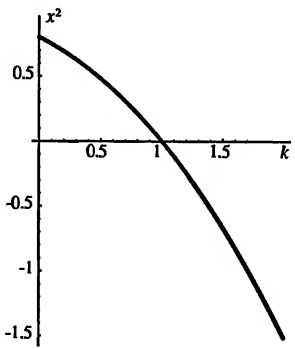


FIGURE 10.3. This is part of the graph of k versus x^2, where for any k, $\pm x$ are the x-coordinates of the intersection points in Figure 10.2. In this graph, $k < 1 \Rightarrow x^2 > 0$, so the intersection points have real x-coordinates. But $k > 1 \Rightarrow x^2 < 0$, and in this case the intersection points have imaginary x-coordinates. In this regard, the intersection points behave like those in the rising parabola in Figure 5.3 on p. 71.

Philosopher. This is good news, and more evidence for my conjecture!

Student. That's nice, but I'd like to see a bigger, more complete picture.

Teacher. Meaning?

Student. I'd like to see some 3-D slice that reveals more of what's actually happening.

Teacher. A good idea. In fact, a short sequence of such pictures might teach us something about how those intersection points move. Since they apparently can have imaginary x-coordinates, it seems reasonable to look in (x_1, ix_2, y_1)-space. Here's that slice of the two ellipses, before any translation:

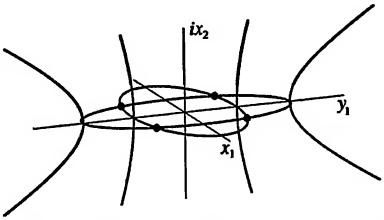


FIGURE 10.4. The two ellipses in Figure 10.1 appear here in the higher-dimensional viewing screen, (x_1, ix_2, y_1)-space. All four intersection points of course appear in the (x_1, y_1)-plane. Notice that the hyperbola branches touching the ellipse

$$\frac{x^2}{1^2} + \frac{y^2}{2^2} = 1$$

are stretched out in the y_1-direction, while the branches associated with the other ellipse are more blunt. This means that no branches intersect.

Now in Figure 10.1, let's push the ellipse

$$\frac{x^2}{1^2} + \frac{y^2}{2^2} = 1$$

upward. The two intersections on the left coalesce, then disappear from the (x_1, y_1)-plane. Where do they go? The next picture tells us:

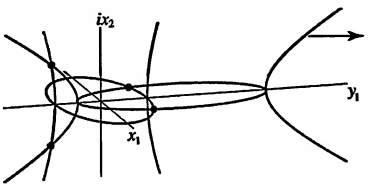


FIGURE 10.5. Here's a picture after the ellipse

$$\frac{x^2}{1^2} + \frac{y^2}{2^2} = 1$$

in Figure 10.1 has been pushed upward—actually, pushed to the right in this 3-D slice. With continued pushing, after two intersection points coalesce they travel along the left hyperbolic branch of the stationary ellipse. As the ellipse continues moving to the right, its left branch intersects the more blunt left branch of the fixed ellipse. But the remaining two intersection points in the (x_1, y_1)-plane are themselves now getting close together, and...

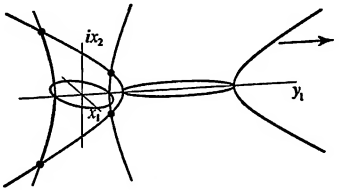


FIGURE 10.6. ...after they've coalesced, the left hyperbolic branch of the moving ellipse intersects not only the left, but now also the *right* hyperbolic branch of the fixed ellipse. All four intersection points are now complex, and they continue climbing along the fixed ellipse's hyperbolic branches as the moving ellipse continues its rightward journey.

Philosopher. This is inspired! These pictures give me a good feeling for where these intersection points are going. They do indeed mimic what happened in the rising parabola of Figure 5.3 [p. 71]. I assume that if the tall ellipse moved down instead of up (left instead of right in these figures), then the intersection points would eventually all lie on the *right* hyperbolic branch of the moving ellipse. In this case also, the points would ultimately travel up both hyperbola branches of the fixed ellipse.

Animated Applet

For an animated version of Figures 10.4–10.6, see Chapter Applet 8 on the CD-ROM.

Teacher. Yes, and that's yet *more* evidence that two conics intersect in four points. Three cheers for $\mathbb{C}^2$!

> In constructing new theories, physicists take certain conservation laws as fundamental: the theory must obey a conservation law for energy, one for momentum, one for electric charge, along with a few others. In constructing a notion of intersection multiplicity, the fundamental idea is this: the total number of intersection points remains the same when we change the coefficients of the defining equations. We therefore choose the surrounding space, define multiplicity and phrase theorems so that "conserving the number of intersection points" remains true.

Student. Maybe only two cheers, because $\mathbb{C}^2$ isn't perfect, is it? I mean, we had to "improve" $\mathbb{R}^2$ by adding points at infinity so that parallel lines would intersect.

Teacher. You're quite right, and you bring up an important question. If two real parallel lines don't meet in $\mathbb{R}^2$, the complex lines containing them don't meet in $\mathbb{C}^2$, either.

Philosopher. Can you prove that?

Teacher. Two parallel lines in $\mathbb{R}^2$ have equations $ax + by = c$ and $ax + by = d$, where $c \neq d$. But the left sides of these two line equations are the same, so if they intersect in $\mathbb{R}^2$, we get a contradiction: $c = d$. The only way out is that the two equations have no finite solution—there is no finite intersection point. In $\mathbb{C}^2$, we can similarly consider two lines having equations $ax + by = c$ and $ax + by = d$, where now the quantities can be complex. We can algebraically define two complex lines to be *parallel* if they have such equations, with $c \neq d$. Then we run up against exactly the same contradiction: the lines cannot intersect in $\mathbb{C}^2$.

Philosopher. So my idea about degenerate conics intersecting in four points isn't quite right, is it? If the lines in the conics are all parallel, we don't get *any* points of intersection, and this seems to say that we must add points at infinity to $\mathbb{C}^2$, too! Do you agree?

Teacher. There seems to be no way out—so yes, I agree.

Is a Circle Really Round?

Student. I think we can run into trouble even when both conics are nondegenerate. I'm thinking about two concentric circles—sort of "parallel circles." I don't think they'd intersect, *ever*!

Teacher. I agree that two circles like $x^2 + y^2 = 1$ and $x^2 + y^2 = 2$ can't have a common solution in $\mathbb{C}^2$, because if they did—that is, if the equations were consistent—then by equating left-hand sides, we'd get $1 = 2$, a contradiction.

Philosopher. So they must intersect at infinity, then. And I hope there are four intersection points there!

Student. I bet there *aren't*! Even at infinity, the intersections would be at some four *specific* points. And then one could ask why not some *other* four points? With the complete symmetry of a circle, there shouldn't be some specific four points. I think two concentric circles never intersect, even at infinity!

Philosopher. Yes! You're starting to use symmetry!

Teacher. But a circle *isn't* symmetric!

Philosopher and Student (Silence, thinking *Teacher* has taken leave of his senses).

Philosopher. You're not referring to a real circle, are you? You're saying a *complex* circle isn't symmetric?

> If you could see really well in four dimensions, then by looking at the complex circle $x^2 + y^2 = 1$ in $\mathbb{C}^2$, you could locate the x- and y-axes. The complex circle in $\mathbb{C}^2 = \mathbb{R}^4$ is not "round" there. In fact, you can take an ordinary rotation of the real circle about its center and extend it to a rotation of $\mathbb{R}^4$ so that the entire complex circle very obviously moves.

Teacher. Yes, that's what I mean. Look at things this way: the equation $x + y = 1$ is symmetric, yet it defines a line with a specific slope. Likewise, $x^4 + y^4 = 1$ is symmetric

in x and y, and so is $|x|^{1/2} + |y|^{1/2} = 1$. But all of these curves are symmetric only in that you can interchange x and y without changing their equations. That means, geometrically, that the curves are symmetric with respect to the lines $y = \pm x$. There's no guarantee of rotational symmetry.

Student. You're saying that the curves defined by your last two equations aren't round...

Teacher. That's right. Take a look:

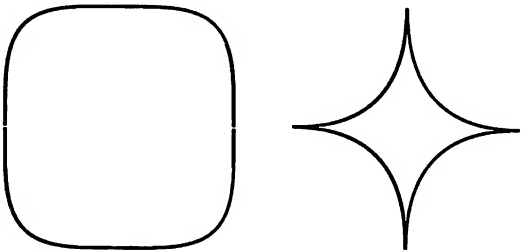


FIGURE 10.7. Although a complex circle doesn't possess rotational symmetry in $\mathbb{R}^4$, we don't see that fact in the usual real plane. The complex curves $x^4 + y^4 = 1$ and $|x|^{1/2} + |y|^{1/2} = 1$ similarly lack such rotational symmetry in $\mathbb{R}^4$. The difference is that this lack spills over into $\mathbb{R}^2$, where it's obvious.

Symmetry is a double-edged sword—it can inspire, but if one isn't careful, it can also mislead.

Philosopher. I like to be open to nontraditional ideas, but I'm afraid you're cracking some of my cherished assumptions about circles!

Teacher. I think this points to what we've been discovering all along: looking at polynomial loci in the real setting reveals such a tiny amount of the total picture, that we can easily end up with wrong ideas.

Philosopher. I shouldn't complain! I really want four intersection points at infinity, and if the complex circle isn't symmetric, and I get my four points, and my conjecture proves true...

Teacher. Whatever the truth turns out to be, it will probably be very beautiful.

Student. What a trip!

Constructing A Beautiful Universe

Teacher. We've discovered that we must add points at infinity to $\mathbb{C}^2$. The burning question is: *how*?

Philosopher. I think we should proceed in analogy with the real case.

> In deciding how to add points at infinity to $\mathbb{C}^2$, the touchstone is that any two distinct complex lines intersect in one point. No matter how the distinct lines are placed, we want the extension of $\mathbb{C}^2$ to contain exactly one intersection point. This is reminiscent of the physicist insisting, in a new theory, that basic conservation laws be preserved.

And I believe that in extending any plane, the basic requirement should be that any two distinct lines should intersect in one point.

Teacher. That sounds reasonable.

Philosopher. In the real setting, the essence of the idea of going from $\mathbb{R}^2$ to $\mathbb{P}^2(\mathbb{R})$ is to add one point at infinity to each line in such a way that parallel lines meet at that added point.

Teacher. Right. Now Definition 1.1 [p. 14] defines $\mathbb{P}^2(\mathbb{R})$ as the set of 1-spaces of $\mathbb{R}^3$, and this approach is quite versatile. In our case, we simply replace $\mathbb{R}$ by $\mathbb{C}$.

Definition 10.1. Complex projective 2-space, denoted $\mathbb{P}^2(\mathbb{C})$, consists of the complex 1-spaces of $\mathbb{C}^3$.

The tremendous power of the replacement idea shows itself in this sweeping "meta-statement":

In Definitions 1.1 through 2.3 [pp. 14–35], we can replace $\mathbb{R}$ by $\mathbb{C}$ to get complex analogues.

In addition, one can simply rewrite all the examples illustrating these definitions, using $\mathbb{C}$ in place of $\mathbb{R}$. The linear algebra statement is always one dimension higher than the corresponding projective statement. In this case, "dimension" of course means complex dimension. Here's a quick summary of some main points:

- **The (complex projective) lines of $\mathbb{P}^2(\mathbb{C})$** are the complex 2-spaces of $\mathbb{C}^3$.
- We can single out one complex projective line in $\mathbb{P}^2(\mathbb{C})$, calling it **the (complex projective) line at infinity**, denoted by $\mathbb{P}^1_\infty(\mathbb{C})$. Relative to this choice, the points of $\mathbb{P}^1_\infty(\mathbb{C})$ are then the **points at infinity of $\mathbb{P}^2(\mathbb{C})$**. Any two distinct complex lines in $\mathbb{P}^2(\mathbb{C})$ intersect in a single point.
- The group of nonsingular linear transformations on $\mathbb{C}^3$ induces on $\mathbb{P}^2(\mathbb{C})$ the **group of projective transformations on $\mathbb{P}^2(\mathbb{C})$**. The study of properties in $\mathbb{P}^2(\mathbb{C})$ invariant under this group of projective transformations constitutes **complex projective plane geometry.**
- The **(complex) homogenization** of a subset S of $\mathbb{C}^3$ consists of the union of all complex 1-spaces of $\mathbb{C}^3$ through points of S.
- The **natural topology on $\mathbb{P}^2(\mathbb{C})$** has as its open sets the complex homogenizations of the open sets of $\mathbb{C}^3$.
- The **closed homogenization** of a subset S of $\mathbb{C}^3$ is the closure in $\mathbb{C}^3$ of the ordinary complex homogenization of S.

- The notions of **homogeneous polynomial** and of **homogenizing** and **dehomogenizing** polynomials are exactly the same as in Definitions 2.9–2.10, with polynomials now having coefficients in $\mathbb{C}$.
- Topologically, the line at infinity of $\mathbb{C}^2$ is a complex line which has been completed by adding a point at infinity to form $\mathbb{P}^1(\mathbb{C})$.

Philosopher. Let me try to recap and get some perspective on all this. We started using $\mathbb{C}^2$ to give two ellipses at least a chance of intersecting in four points, and the evidence showed that $\mathbb{C}^2$ has to be part of the story. But by looking at degenerate parallel-line examples, we saw that $\mathbb{C}^2$ wasn't the last word, so we added points at infinity to $\mathbb{C}^2$ to get $\mathbb{P}^2(\mathbb{C})$, in which such lines do intersect. The question is, are we about to encounter further difficulties, or is $\mathbb{P}^2(\mathbb{C})$ really the right universe to work in?

Since $\mathbb{P}^1(\mathbb{R})$ is a topological circle, it's orientable. On the other hand, we know that $\mathbb{P}^2(\mathbb{R})$ is not orientable. The general situation is this: $\mathbb{P}^n(\mathbb{R})$—which can be defined as the set of all 1-subspaces of $\mathbb{R}^{n+1}$—is orientable if n is odd, and nonorientable if n is even.

For the sophisticate

The above definitions of real or complex projective space generalize: let k be any field equipped with a topology, and use that topology to define the product topology on the vector space k^{n+1}, $n \geq 0$. Then define $\mathbb{P}^n(k)$ to be the set of all 1-spaces of k^{n+1}, and define an open set of $\mathbb{P}^n(k)$ to consist of all 1-spaces in k^{n+1} intersecting some open set of k^{n+1}. Let k^n be some fixed n-dimensional subspace of k^{n+1}, and let $\mathbf{v}$ be any vector in k^{n+1} not in k^n. For each point of $\mathbf{v} + k^n$, there is a corresponding 1-subspace of k^{n+1} containing that point; this 1-space is not in k^n. We can regard k^{n+1} as a cone, and $\mathbf{v} + k^n$ as a dehomogenization of that cone relative to our choice of k^n, this dehomogenization being a copy of affine n-space. The 1-spaces of k^n, being all the 1-spaces of k^{n+1} that don't intersect $\mathbf{v} + k^n$, are therefore the points at infinity of $\mathbf{v} + k^n$ Those 1-spaces of k^n, by our definition, form a projective space $\mathbb{P}^{n-1}(k)$. That is to say, *the points at infinity of any projective space* $\mathbb{P}^n(k)$ *form a projective space* $\mathbb{P}^{n-1}(k)$.

Teacher. The news is good. $\mathbb{P}^2(\mathbb{C})$ is an excellent universe to work in. In fact, your conjecture is an instance of one of the most powerful and beautiful results in mathematics—Bézout's theorem, and $\mathbb{P}^2(\mathbb{C})$ is a perfect place in which to state it.

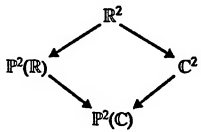
Student. Wow—*one of the most powerful and beautiful results in mathematics*! You don't talk like that too often.

Teacher. One doesn't meet theorems like this too often.

Philosopher. It sounds like you're leading up to something big! Exactly what *is* Bézout's theorem?

Chapter Summary

The sequence of pictures in Figures 10.4–10.6 beginning on p. 174 illustrates one great advantage of working in $\mathbb{C}^2$: it gives us a basic "conservation law." In this example, points in which two ellipses intersect don't simply disappear as one ellipse is translated indefinitely—we can forever track all four intersection points. However, $\mathbb{C}^2$ doesn't always save the day—for example, when we rotate one of two intersecting lines, the point of intersection does actually disappear when the lines become parallel. In $\mathbb{C}^2$, as in the real plane, the secret lies in adding one "point at infinity" to each line. Doing this leads to $\mathbb{P}^2(\mathbb{C})$, and this turns out to be the ideal "universe" in which to view conics and other polynomially-defined curves, track intersection points and state (as well as prove) various facts. The stages we've encountered in our journey of extending the real plane can be summed up in the diagram



Although it's not easy to visualize the transitions $\mathbb{P}^2(\mathbb{R}) \to \mathbb{P}^2(\mathbb{C})$ or $\mathbb{C}^2 \to \mathbb{P}^2(\mathbb{C})$, the idea of defining $\mathbb{P}^2(\mathbb{R})$ as the set of 1-spaces of $\mathbb{R}^3$ generalizes directly to the complex setting, and this simple algebraic definition makes it easy to establish various basic facts. A sampling of such facts appears after Definition 10.1.

Exercises

1. One can use software such as Maple or Mathematica to translate those intersections in Figures 10.4–10.6 into specific coordinatized points in $\mathbb{C}^2$. For example, suppose that the moving ellipse went 0.5 to the right in going from Figure 10.4 to Figure 10.5. Then Maple can find the four intersection points by executing the command

$$\textbf{solve}(\{x\hat{}2 + 4*y\hat{}2 = 4.0, 4*(x - 0.5)\hat{}2 + y\hat{}2 = 4.0\});$$

The corresponding command in Mathematica is

$$\textbf{Solve}[\{x\hat{}2 + 4*y\hat{}2 == 4.0, 4*(x - 0.5)\hat{}2 + y\hat{}2 == 4.0\}, \{x, y\}]$$

These return two real points and two complex points, as the figure suggests. (We've cleared denominators for convenience. Also, introducing a decimal like 4.0 generally produces answers in decimal form.) *Exercise*: Run a series of such computations showing that what the figures suggest actually does happen: the initial four real, distinct intersection points transition to two real points and one double point, then two real and two complex points, and so on.

2. Which of these are concepts belonging to two-dimensional complex projective geometry?
   - Principal complex axes of a complex ellipse;
   - Orthogonality of those principal axes;
   - "Inside" versus "outside" of a nondegenerate conic;
   - Degeneracy of a conic;
   - Tangent line to a conic;
   - Two conics intersecting transversally (that is, not tangentially) at a point.
3. Just as one can define vector spaces over any field of scalars, one can define projective spaces over any field k, and if the field is finite, one can work out various instructive "counting rules." For example, one can let k be the "five-hour clock" with hours 0, 1, 2, 3, 4 (0 playing the role of the "12" on a standard 12-hour clock), equipped with addition and multiplication modulo 5. (Example: $2 \div 3 = \frac{2}{3}$ can be evaluated by writing $\frac{2}{3} = x$. This gives $2 = 3x$, which in turn has the unique solution $x = 4$.)
   (a) Draw the vector space k^2.
   (b) How many 1-spaces are there of k^2? How many points in each 1-space? Write down the coordinates of the points in each 1-space.
   (c) *Complete* each family of parallel lines by adding a common "point at infinity" to each of its lines. Call the set L_∞ of added points *the line at infinity*. Does L_∞ have as many points as each of your completed lines?
   (d) Denote $k^2 \cup L_\infty$ by $\mathbb{P}^2(k)$. Show that the number of points in $\mathbb{P}^2(k)$ coincides with the number of 1-subspaces of k^3 (Hint: most 1-spaces of k^3, intersect the plane $x = 1$ there in one point.)
4. (a) In the k^2 of Exercise 3, draw a circle $x^2 + y^2 = 1$.
   (b) In k^2, how many circles are there of nonzero radius, centered at the origin?
   (c) Are any two of these concentric circles disjoint?
   (d) List the points in k^2 comprising some noncircle "ellipse"

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

   in k^2.
   (e) List the points in k^2 comprising some "hyperbola"

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$$

   in k^2.
   (f) For what radius (or radii) r is it most likely that two circles of radius r will intersect? (Assume they have different centers.)
5. It will follow from what we'll learn in the next chapter, that if two complex circles intersect in more than four points, then they must actually be the same circle. Yet the box on p. 176 says that if you regard a complex circle as a subset of $\mathbb{R}^4$, you can give it a Euclidean rotation about its center that changes it to a different set although, just

looking in the real plane, the rotation simply sends the circle to itself. That means the original complex circle intersects its rotated image in all those infinitely many real points. Is there a contradiction here? Explain.

6. In analogy to $\mathbb{P}^1(\mathbb{R})$ and $\mathbb{P}^2(\mathbb{R})$, the three canonical affine dehomogenizations of $\mathbb{P}^2(\mathbb{C})$ form an open cover of it, each set of the cover therefore being a copy of $\mathbb{R}^4$. Find real coordinate maps between each pair of these open sets so that the Jacobian of the map has positive determinant at each point. This is the key step in one proof that, as a real 4-manifold, $\mathbb{P}^2(\mathbb{C})$ is orientable. (See the end of Appendix 2.)

11

A Most Excellent Theorem

in which the universe just constructed pays off!

The "Atoms" of Our Theorem

Teacher. In days long past, stories occasionally surfaced about somebody discovering an old violin in an attic which would turn out to be a Stradivarius—a find that would enrich both the finder and the musical world. Well, Bézout's theorem is something like that—you accidentally run across it, and if you're in the right frame of mind you examine it more closely, finally realizing that you've stumbled across an incredible treasure! The fact is, most people haven't heard of it—unlike the fundamental theorem of algebra, for example. Yet Bézout's theorem is actually more general and more beautiful.

What he proved is a theorem saying how often curves defined by polynomials can intersect. Let's begin with a few basic definitions and facts about polynomials and the curves they define. First, each set of polynomials $\mathbb{C}[x]$, $\mathbb{C}[x, y]$, $\mathbb{C}[x, y, z]$, ... is a "unique factorization domain." This means that each nonconstant polynomial in any of $\mathbb{C}[x]$, $\mathbb{C}[x, y]$, $\mathbb{C}[x, y, z]$, ... can be factored into a product of *irreducible polynomials*—ones that are not the product of two nonconstant polynomials. Now a *unit* in any of $\mathbb{C}[x]$, $\mathbb{C}[x, y]$, $\mathbb{C}[x, y, z]$, ... is an element having a multiplicative inverse; these are just the nonzero elements of $\mathbb{C}$. It turns out that the decomposition of a polynomial into a product of irreducible factors is unique up to units and order of the factors.

Student. That sure sounds like something I learned in a number theory course: every integer is a product of primes! Prime numbers are defined in a similar way and, except for the order of the factors, the product is unique up to multiplying them by ± 1.

Teacher. That's right. And ± 1 are the only integers having integer inverses, so they're the units of $\mathbb{Z}$. Prime numbers and irreducible polynomials are like "atoms."

Philosopher. "Atoms of construction"? I get the feeling that we're looking at two peaks of a submerged mountain range—the Hawaiian Islands again!

Teacher. I agree, and there are further peaks. It turns out that factoring polynomials has a direct geometric translation. We illustrate in $\mathbb{C}[x, y]$, though the same idea works in any number of variables. For $p \in \mathbb{C}[x, y]$, let $C(p)$ denote the locus of p in $\mathbb{C}^2$—that is, the set

of points $(a, b) \in \mathbb{C}^2$ for which $p(a, b) = 0$. The letter "C" stands for "curve": the locus in $\mathbb{C}^2$ has complex dimension 1, and $\mathcal{C}(p)$ is called an *algebraic curve in* $\mathbb{C}^2$.

It turns out that we can transport most of the vital stats of polynomials to vital stats of algebraic curves. For example, an algebraic curve $C(p)$ is defined to be *irreducible* if p is irreducible, and any algebraic curve $C(p)$ is the union of irreducible algebraic curves. That is, if p factors into irreducibles as $p = p_1 p_2 \cdots p_n$, then $C(p) = C(p_1) \cup \cdots \cup C(p_n)$.

Student. You could just write that last as $C(p_1 \cdot p_2 \cdots p_n) = C(p_1) \cup \cdots \cup C(p_n)$.

Teacher. Yes, and here's why we have equality:

- $C(p_1) \cup \cdots \cup C(p_n) \subseteq C(p_1 p_2 \cdots p_n)$: Any point P in $C(p_1) \cup \cdots \cup C(p_n)$ must be in some $C(p_i)$. That means $p_i(P)$ is 0, which of course means that the product $p(P) = p_1(P)p_2(P) \cdots p_n(P)$ is also 0. Therefore, P is in $C(p_1 \cdot p_2 \cdots p_n)$.
- $C(p_1 p_2 \cdots p_n) \subseteq C(p_1) \cup \cdots \cup C(p_n)$: If a point P is in $C(p_1 p_2 \cdots p_n)$, then $p_1(P)p_2(P) \cdots p_n(P) = 0$. Now there are no zero-divisors in $\mathbb{R}$ or $\mathbb{C}$, so some $p_i(P)$ must be zero. This means that P is in at least one of $C(p_1), \ldots, C(p_n)$, so of course it's in $C(p_1) \cup \cdots \cup C(p_n)$.

Student. I always wondered if there was some one equation for the union of *two* circles, and you seem to be saying that there is.

Teacher. Indeed there is. Take two circles—say of radii 1 and 2 and centered at the origin. Their equations are $x^2 + y^2 = 1$ and $x^2 + y^2 = 4$, and we can write these as

$$x^2 + y^2 - 1 = 0, \quad x^2 + y^2 - 4 = 0.$$

The equation for their union is therefore $(x^2 + y^2 - 1)(x^2 + y^2 - 4) = 0$. To amaze your friends, you could cover your steps by multiplying everything out, arriving at an unlikely equation for the union of our two circles, such as

$$x^4 + y^4 + 2x^2y^2 + 4 = 5(x^2 + y^2).$$

Student. That is really *great*!

Teacher. Getting back to our factorization of p, just as with integers, factors p_i may be repeated, but if p_i and p_j differ by more than just a unit multiple, then $C(p_i) \neq C(p_j)$. Those $C(p_i)$ are called the *irreducible components of* $C(p)$. For convenience, we'll always assume that unless stated otherwise, there are no repeated factors. One can then easily transport the notion of degree of a curve: the *degree* of $C(p)$ is defined to be the degree of p, and in obvious notation, one has

$$\deg(p) = \deg(p_1) + \cdots + \deg(p_n) \Rightarrow \deg C(p) = \deg C(p_1) + \cdots + \deg C(p_n).$$

Philosopher. This translating between algebra and geometry reminds me somewhat of what one sees in analytic geometry.

Teacher. I agree. These are basic notions from "plane affine algebraic geometry"—algebraic geometry in $\mathbb{C}^2$ or $\mathbb{R}^2$. To some extent, two-dimensional analytic geometry is algebraic geometry in $\mathbb{R}^2$, but restricted to the geometric world of points, lines, circles and their algebraic counterparts.

Remarkably, we can carry these affine ideas to the projective level, the curves then living in the beautiful world $\mathbb{P}^2(\mathbb{C})$ that we considered in the last chapter. It is in this world that

Bézout's theorem holds. Here are the corresponding ideas: any polynomial $p(x, y)$ can be homogenized to a polynomial $H(p)(x, y, z)$ as described in Definition 2.1 [p. 33], and it is an easy exercise to show that a homogeneous polynomial factors into homogeneous irreducibles. The locus of a homogeneous polynomial is a homogeneous set in $\mathbb{C}^3$ which therefore defines a set in $\mathbb{P}^2(\mathbb{C})$ [Definition 10.1, p. 178]. We variously refer to these sets by terms such as *projective algebraic curves, algebraic curves in* $\mathbb{P}^2(\mathbb{C})$, or *plane projective curves.* An algebraic curve in $\mathbb{P}^2(\mathbb{C})$ is *irreducible* if a homogeneous polynomial defining it is irreducible, and one has, just as above,

$$H(p) = H(p_1)H(p_2)\cdots H(p_n) \Rightarrow C(H(p)) = C(H(p_1)) \cup \cdots \cup C(H(p_n)).$$

Again, factors of $H(p)$ may be repeated, but if $H(p_i)$ and $H(p_j)$ are irreducible and do not differ by just a unit multiple, then $C(H(p_i)) \neq C(H(p_j))$. Not surprisingly, the $C(H(p_i))$ are called the *irreducible components of* $C(H(p))$. The degree of a polynomial is unaffected by homogenizing, and one can transfer "degree of an algebraic curve" to the projective setting, obtaining:

$$\deg(H(p)) = \deg(H(p_1)) + \cdots + \deg(H(p_n))$$
$$\Rightarrow \deg C(H(p)) = \deg C(H(p_1)) + \cdots + \deg C(H(p_n)).$$

Student. So when are we going to find out what Bézout's theorem is?!

Bézout's Theorem

Teacher. Here goes. Bézout's theorem roughly says this:

Bézout's theorem. Let C_1 and C_2 be algebraic curves in $\mathbb{P}^2(\mathbb{C})$ of degrees m and n. Then
either:
C_1 and C_2 intersect in exactly mn points, counted with multiplicities;
or:
C_1 and C_2 share at least one irreducible component.

The basic idea of this theorem is that two algebraic curves C_1 and C_2 almost always intersect in $\deg(C_1) \cdot \deg(C_2)$ points. The only exception is when the intersection contains a curve—that is, when an entire component of C_1 coincides with a component of C_2.

Philosopher. So if C_1 and C_2 are two different ellipses, then because their polynomial equations are both quadratic, the ellipses ought to intersect in $2 \cdot 2 = 4$ points?

Teacher. That's right, and that's why your conjecture about four points has to be correct.

Student. You're saying that if I don't see four points, it's because I just haven't looked hard enough?

Teacher. Exactly. One obvious example is that some intersection points, or even all of them, may lie outside the familiar real viewing screen, their coordinates being complex. Another possibility, as we've noted, is that some points may be "stuck together," looking like one point, when in fact they're several points lying on top of one another. Of course we need to learn how to get them "unstuck" so we can count them correctly. And relative to looking in $\mathbb{R}^2$ or $\mathbb{C}^2$, yet another hiding place is the line at infinity. But Bézout's theorem tells us that if we do a complete accounting, we get precisely mn points.

Philosopher. So if I intersect an ellipse and a cubic, I must get *exactly six* intersection points, and if I find only five, I can confidently continue my search, knowing that there's got to be another one out there!

Teacher. That's it.

Student. Suppose I find *seven* points?

Teacher. That can happen, but then the second part of the theorem says that the ellipse (which happens to be irreducible) must be part of the cubic.

Student. How could an ellipse ever be part of a cubic? I've never seen a cubic that looks like *that*.

Teacher. The union of an ellipse and a line is a cubic that looks that way. Consider, for example, the circle $x^2 + y^2 - 1 = 0$ and the x-axis $y = 0$. Their union is the solution set of the product $(x^2 - y^2 - 1)y$ set equal to zero, and this polynomial has degree 3.

Student. That Bézout theorem really *is* powerful!

Teacher. What did I tell you? Now I used the word *roughly*, because the phrase "counted with multiplicities" conceals some fine print in the full "airtight" statement of the theorem.

Étienne Bézout 1730–1783

His famous theorem was first published in 1779. When Étienne was a youngster, it was assumed he would follow in his father's footsteps as a governmental official, but all that changed when the talented boy read some of Euler's works and fell under their spell. Later in life, Bézout was much admired for his clear, understandable textbooks, and was known for his warm heart and kindness.

Student. So how *do* you count with multiplicity? What does that mean?

Teacher. Well, as an illustration, we saw that when two ellipses are tangent, it's natural to count the point of tangency with multiplicity 2, since two points coalesced there, and we can think of the intersection point as actually one point on top of another. But things can get more subtle than that. For example, one or both curves may have self-intersection

points, looking right around such a point like an "X," and that increases the multiplicity if the curves intersect at such a point. Even intersections at infinity need to be counted with multiplicities.

Student. Maybe that fine print is why Bézout's theorem isn't so well known!

Teacher. Perhaps. But a few well-chosen examples can give one a pretty good sense of what the fine print amounts to. Everything is actually quite natural and sensible.

Student. But I still have a problem. We're studying conics, which have degree 2. You're now introducing this general theorem with all the fine print. I remember a teacher once told me not to use a cannon if a fly swatter will do!

Teacher. You raise a good point, and it deserves a good answer. To me, it seems that a general theorem puts us on a mountainside, allowing us to survey a panorama. A more general result puts us higher up on the mountain, where we can see an even vaster area, and such a view can be very satisfying in its own right. Now, we could be completely utilitarian about things, and prove no more than what's needed at the moment—essentially Bézout's theorem for $n = m = 2$. At that point, we might even wonder whether the right way to look at things is as $4 = 2 \cdot 2$, or whether $4 = 2 + 2$ is the right way. Bézout's theorem, as we state it, tells us exactly what the spirit is. Besides, not all facts about conics are degree-two in nature; Bézout's theorem allows us to establish a number of results of a higher-degree nature.

Student. *Really*? I always assumed studying conics was a "closed world," where everything has degree 2.

Philosopher. Do I sense another reference to the "Hawaiian Islands principle" coming?

Teacher. Yes. Suppose you inscribe a hexagon in an ellipse. If you extend the pairs of opposite sides of an inscribed hexagon, then these three pairs intersect in three points...

Student (Interrupting). But suppose you inscribe a regular hexagon in a circle? Then opposite sides don't intersect at all!

Teacher. They intersect at infinity in three points, because we assume we've added in the line at infinity. Now Blaise Pascal (1623–1662), as a teenager, noticed a remarkable phenomenon: the three points always lie on a straight line. This fact—"Pascal's theorem"—has a nice proof using Bézout's theorem, and in that proof, one encounters degree three. I will show you a proof in a while.

Student. You make a pretty convincing case. I know I interrupted you. You said a minute ago that you had some examples of Bézout's theorem. I like examples, and find it easy to learn from them, so I'm ready to see some, especially if there are pictures!

Rare Events

Teacher. OK, let's begin with that picture of the rising parabola—Figure 5.3 [p. 71]. Notice that we could just as well keep the parabola fixed and move the horizontal axis up or down. In only one position, however, does the horizontal line intersect the parabola in a double point, and that's when the line is tangent to the parabola. From that position, if you give the line just the slightest nudge up or down, then it loses the special property of being tangent. The two coalesced points—a point of multiplicity 2—then separate out into two

single points. In a sense, tangency is rare: like a pencil precariously balanced on its point, the slightest nudge brings the house of cards tumbling down to a more stable state. In this case, a stable state is when a line intersects the parabola nontangentially at a point—that is, *transversally* there. If you nudge such a line a little this way or that, you still get one point of transverse intersection within the immediate neighborhood. As Figure 5.3 suggests, the points may be complex, but to someone with "4-D vision," all this is very evident.

Philosopher. Your analogy makes me think of artificial elements numbered over 100, say. They're so heavy that they can manage to stay together for only a small fraction of a second. *Their* house of cards then comes crashing down, the element decaying into several lighter, more stable particles.

Teacher. An interesting analogy. In general, an intersection point of multiplicity > 1 is in a sense "rare." If we just shake things up a bit, it typically splits apart into several points, each having multiplicity 1. To find the original multiplicity, just count the points.

Philosopher. What constitutes "shaking things up a bit"?

Teacher. Just about anything you do to the curve or its equation: translate the curve a little in some direction, rotate it a bit around some point, or slightly change one or more of the polynomial's coefficients. In the complex setting, such small perturbations will nearly always knock an intersection point of higher multiplicity from its precarious state, leaving instead several intersection points, each having multiplicity 1.

> "Perturbing" an equation almost always splits up any multiple points in its locus, but in the usual real setting we don't always see much evidence of this. But by choosing just the right perturbation—usually a combination of two or three basic ones such as a translation, rotation or slightly changing one of the defining polynomial's coefficients—we can usually coax the equation to separate all of its multiple points in the real setting.

Student. This sounds exciting, but I need to see some pictures!

Teacher. It's time for that. Instead of the parabola $y = x^2$, let's look at the cubic $y = x^3$. The x-axis is again tangent to the curve, and that means that the multiplicity of their intersection point (0, 0) is greater than 1. When we move the x-axis up or down a little, the point of tangency splits up into *three* points, not just two.

Student. Well, I certainly don't see anything splitting up—the horizontal line always intersects the curve in one point.

Teacher. That's exactly the problem with looking only at the real part. If we could see in four dimensions, we'd see all three points.

Student. But we see only one of them, so what now?

Teacher. We have a couple of options. For one, we could algebraically solve for the intersection points, and that would yield three solutions, corresponding to three points. Another option is to try perturbing in *other* ways—not just translating—to tease them apart in the real plane so we can see the separation take place.

Philosopher. How does the algebraic approach work?

Teacher. The complex x-axis, translated upward by a, has equation $y = a$. To find the points of intersection of this translated x-axis with the complex graph of $y = x^3$, equate the y's. This gives $x^3 = a$, so the x-coordinates of the intersection points are the three cube roots of a. These roots are sketched in the right picture of Figure 11.1 in the cases of a positive (solid lines) and a negative (dashed lines):

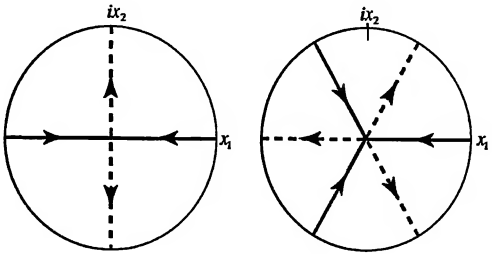


FIGURE 11.1. The picture on the left shows the paths of solutions of $a = x^2$ in the complex x-axis, as a goes from positive to negative. The picture on the right shows corresponding solutions for $a = x^3$. In general, solutions to $a = x^n$ approach the origin along the n rays

$$\theta = \frac{2\pi j}{n} \qquad (j = 0, 1, \ldots, n-1),$$

and they separate along the angle bisectors of adjacent incoming paths.

As the complex x-axis is moved downward from $a > 0$ to 0, the three x-values approach the origin along the solid lines in the figure on the right. The cubic's ordinary *real* graph of course lives in the (x_1, y_1)-plane. Now in (x_1, ix_2, y_1)-space, rotate that plane about the y_1-axis by 120° and 240°. The union of the original graph and its two rotated copies forms the part of the complex graph of $y = x^3$ contained in (x_1, ix_2, y_1)-space. So as we translate the complex x-axis toward the origin in (x_1, ix_2, y_1)-space, we see three points approaching the origin and coalescing there. When we further translate the x-axis downward, the triple point symmetrically splits up into three single points, each traveling along a straight path making a 60° angle with one of the original paths. For $y = x^n$, n points analogously symmetrically approach the origin along the rays defined by

$$\theta = \frac{2j\pi}{n}$$

and separate along rays

$$\theta = \frac{(2j+1)\pi}{n} \qquad (j = 0, \ldots, n-1).$$

Student. OK, I see the points. But you mentioned coaxing the points to separate in the real plane by doing something other than translating. Translating is all we've done so far.

Teacher. True. For the cubic $y = x^3$, we could simply rotate the x-axis a little. Look at what we get:

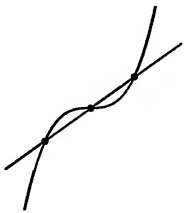


FIGURE 11.2. The x-axis intersects the cubic $y = x^3$ in (0, 0) with multiplicity 3. In this picture, we've rotated the x-axis about the origin, making (0, 0) separate into three points in the real plane.

Student. Brilliant! All three points are in plain view. And I also see that as you "unrotate" the line back to its original horizontal position, the three points all come back together.

Teacher. I admit that's nice, but I have to say that for $y = x^4$, we'd have to work harder to get the points to separate in the real plane. One way would be to jiggle the lower degree coefficients (which are all 0 to start with) to get four closely-spaced values of x for each small, nonzero value of y.

Philosopher. Something worries me. In the geometric approach to assigning multiplicity, maybe you might not be clever enough to see some way of separating all the stuck-together points so you can see them. It seems that the algebraic approach is both safer and more general.

Teacher. It is. The geometric approach can be intuitive and extremely suggestive, but algebraic results are more reliable, and an algebraic formulation more easily leads to a workable definition.

Student. But I want to see more pictures!

Teacher. Why not do both? Let's draw pictures and supplement them with their algebraic counterpart. We'll get a better appreciation of multiplicity that way.

For example, Figure 11.3, next, shows two suggestive pictures: In each picture, the "alpha curve" $y^2 = x^2(x + 1)$ crosses itself at the origin. In the picture on the left, the solid line is transverse to each of the two intersecting "branches" of the curve, so we have two stable situations. Geometrically, this suggests that the intersection multiplicity is 2. In the picture on the right, the solid line is tangent to one branch, and that tangency increases the intersection multiplicity. Translating into algebra verifies what our eyes tell us. In the picture on the left, let $y = mx$ be parametrized by $x = t, y = mt$. Substituting this into $y^2 = x^2(x + 1)$ gives $(mt)^2 = t^2(t + 1)$, which can be written as

$$t^3 + (1 - m^2)t^2 = 0. \tag{11.1}$$

Choosing any particular t gives a point on the line. If t happens to satisfy (11.1), then the point is on the alpha curve, too, meaning we're at an intersection point. If the root t is

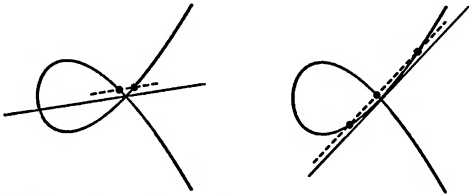


FIGURE 11.3. In each picture a small translation of the solid line to the dashed one splits apart all points in which the line intersects the alpha-curve's double point, or *singularity*. In the curve on the left, the intersection multiplicity at the singularity is 2; in the right picture, the intersection multiplicity at the singularity is 3.

repeated with a certain multiplicity, we define the corresponding point to have that same multiplicity. If $1 - m^2 \neq 0$, then (11.1) shows that $t = 0$ is a double root, and this agrees with what the left picture suggests.

Philosopher. And the picture on the *right*?

Teacher. That corresponds to $1 - m^2 = 0$, where the line is tangent at the origin to one of the branches. It's easy to check algebraically what our eyes tell us, since when $1 - m^2 = 0$, (11.1) becomes $t^3 = 0$. Now **the order of t in a polynomial $p(t)$ (or even in a power series in t) is defined to be the lowest degree of all its terms.** That means the order of t^3 is 3. This means, in turn, that 0 is a root of $t^3 = 0$ of multiplicity 3, and algebraic geometers have shown that this approach, suitably generalized, yields a consistent definition of intersection multiplicity. Therefore we can say that the above line and the alpha curve in the right picture do in fact intersect at (0, 0) with multiplicity 3.

Philosopher. This algebraic approach seems powerful and versatile.

> Counting intersection points with multiplicities is reminiscent of conservation laws in physics. For example, if no electric charges enter or leave a certain region, then no matter what else happens inside the region, the sum total of electric charge remains the same. In our case, consider a region of $\mathbb{C}^2$ in which two curves intersect. No matter how we perturb the coefficients of the curves' equations, as long as no intersection points enter or leave the region, the total sum of the intersection multiplicities stays the same.

Teacher. Yes. It easily works when a line intersects a curve such as $y^2 = x^3$. Take the parametrization $x = t, y = mt$ of a line through (0, 0) and substitute into $y^2 = x^3$. This gives $t^3 - m^2t^2 = 0$, which has order 2 if $m \neq 0$ and order 3 if $m = 0$.

Student. You're always using a line for one of the curves. What if neither curve is a line?

Teacher. If you can find a parametrization of one of the curves, with $t = 0$ corresponding to the intersection point in question, then our method continues to work. As an example,

look at the two curves $y^2 = x^3$ and $x^2 = y^3$. Not surprisingly, these are called "cuspidal cubics" since each of them has a cusp (horn) at the origin.

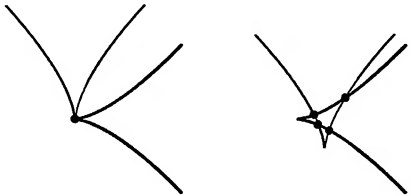


FIGURE 11.4. In the picture on the left, a small translation of one of the curves splits apart a higher-multiplicity intersection point. The picture on the right suggests that this multiplicity of intersection is 4.

They obviously intersect at the origin. What is the intersection multiplicity? We can arrive at the picture on the right by translating one of them a bit, and it suggests that the intersection's multiplicity is 4.

Student. Could there actually be more than four points stuck together at the origin, and then might some other way of perturbing the curves reveal this?

Teacher. That's a possibility we should rule out. Again, doing this would be less of a problem if we could see in four dimensions! But the algebra is remarkably easy: choose one of the curves, say $y^2 = x^3$ and parametrize it by $x = t^2, y = t^3$. Substituting this into the other curve's equation gives $(t^2)^2 = (t^3)^3$—that is, $t^9 - t^4 = 0$. The order is 4, so according to what we said a moment ago, 4 is the intersection multiplicity at the point corresponding to $t = 0$, namely, the origin.

Philosopher. Amazing!

Student. OK, but suppose you have two cuspidal curves that are *tangent*, like this:

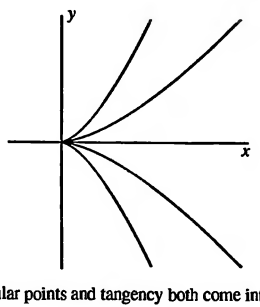


FIGURE 11.5. Singular points and tangency both come into play here. Result: higher multiplicity.

Then what do you do?

Teacher. This is a good example because the geometric approach of separating out the points in the real plane is definitely more challenging. Let's say that the curves' equations are $y^2 = x^3$ and $y^2 = 2x^3$ One could perturb the first equation into $y^2 = x^2(x + \epsilon)$, which turns it into an alpha curve. This in effect "loosens the knot" at the cusp. Doing a similar thing to the other equation, coupled with a little translation, can produce situations like these:

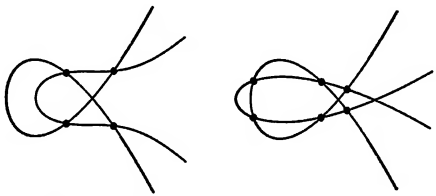


FIGURE 11.6. Figure 11.5 can change into pictures like these when we perturb both curves by appropriately perturbing coefficients and constant terms in their equations. On the left, only four of the six intersection points are real, while on the right, all six are real.

In the right picture, we see six separate points.

Philosopher. But again, is that *actually* the intersection multiplicity? Using the geometric approach is enticing and intuitive, but it seems that using it is as much an art as a science.

Teacher. With our limited ability to see in four dimensions, you're quite right. The intersection multiplicity is in fact 6, and showing this algebraically is a snap: substituting the parametrization $x = t^2$, $y = t^3$ into $y^2 = 2x^3$ gives $(t^3)^2 = 2(t^2)^3$, which simplifies to $t^6 = 0$; the intersection multiplicity of (0, 0) is thus 6.

> Here's a different way of looking at intersection multiplicity: First, imagine a closed curve—a loop—winding n times around a circle to make a donut shape. That winding loop is sometimes called a solenoid, and we say that the solenoid and the circle have *linking number n*. Second, by a small 3-sphere, we mean a translate of $x_1^2 + x_2^2 + y_1^2 + y_2^2 = \epsilon^2$. Now it happens that a sufficiently small 3-sphere centered at a point of a complex curve intersects the curve in a loop. Suppose P is a point of intersection of two complex curves. For ϵ sufficiently small, the 3-sphere centered at P intersects the two complex curves in two linked loops. If you make a string model of these linked loops, then either loop can be massaged into a circle, the other curve then winding around that circle, solenoid-style. *The two loops have linking number equal to the intersection multiplicity at P.*

Interactive Applets

For interactive versions of Figure 11.6, see Chapter Applets 9 and 10 on the CD-ROM.

Finding Missing Points

Student. I'm looking at Figure 11.5 showing two curves having an intersection point of multiplicity 6. There must be additional points where the two cusps intersect, right? The degree of each cusp's equation is 3, so according to Bézout's theorem, there must be nine intersections in all. Where are the other three?

Teacher. In $\mathbb{C}^2$, the origin is the only location common to the two curves. We know this because equating the y's in $y^2 = x^3$ and $y^2 = 2x^3$ gives $x^3 = 2x^3$, which means $0 = x^3$, or $x = 0$. Setting $x = 0$ in either cusp equation implies $y = 0$, meaning (0, 0) is the only solution in $\mathbb{C}^2$ That means...

Student. That means all the other points of intersection must be at infinity!

Teacher. Exactly. Here's a disk model of the two cusps:

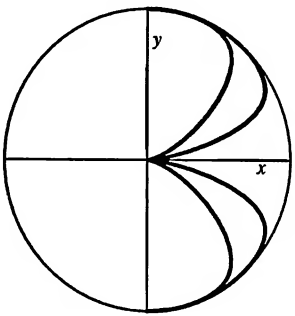


FIGURE 11.7. This shows two tangent cusps in the disk model. Bézout's theorem implies that there must be three intersection points at infinity.

It appears that the curves intersect at the "end" of the y-axis, so let's place ourselves there by the usual homogenization/dehomogenization method. The homogenizations of the curves' equations are $y^2z = x^3$ and $y^2z = 2x^3$. Dehomogenizing at $y = 1$ gives the two cubics $z = x^3$ and $z = 2x^3$, shown on the left in Figure 11.8.

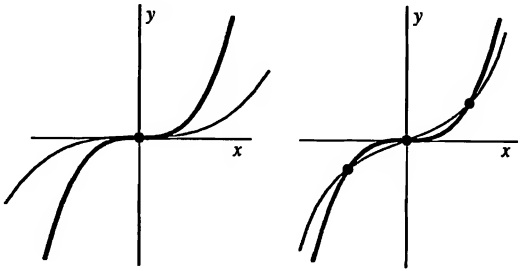


FIGURE 11.8. The picture on the left shows two cubics tangent at their inflection points. Rotating one of the cubics separates three coincident points into three separate ones; the intersection multiplicity on the left is 3.

Rotating one with respect to the other separates the origin into three points, this picture looking a lot like Figure 11.2 [p. 190]. So the intersection point at infinity has at least multiplicity 3. On the other hand, that's exactly the number needed to satisfy Bézout's theorem: $6+3=9$.

> In physics, there are both "local" and "global" conservation laws. The statement "there is always the same total electric charge in a region in which no charge crosses the boundary" is local; while the statement "the net electric charge in the whole universe is always the same" is global. When we completely split up a point with multiplicity and find that we always get the same number of nearby points, that's local. When we choose curves of degree n and m and Bézout's theorem tells us that the total number of their intersection points in the whole "universe" of $\mathbb{P}^2(\mathbb{C})$, counted with multiplicity, must be the constant nm, that's a global conservation theorem.

Philosopher. It would be nice to check that our algebraic method does indeed give multiplicity 3 for the intersection at infinity.

Teacher. A parametrization of $z=x^3$ is $x=t, z=t^3$, and substituting this into $z=2x^3$ gives $t^3=2t^3$, or $t^3=0$. The order of t^3 is indeed 3.

Student. But how about the parabola $y=x^2$ and a vertical translate of it, like $y=x^2+1$? They couldn't have any finite intersections, because if we try to find them by equating y's, we get $x=x+1$, or $0=1$. Yet each equation has degree 2, so Bézout's theorem says there must be *four* points of intersection!

Teacher. Sure—and you know where they're gonna be!

Student. At infinity, again.

Teacher. That's right, and they'll be at the end of the y-axis. So let's homogenize, getting $yz = x^2$ and $yz = x^2 + z^2$. Dehomogenizing at $y = 1$ gives $z = x^2$ and $z = x^2 + z^2$. A parametrization of $z = x^2$ is $x = t$, $z = t^2$, and substituting this into $z = x^2 + z^2$ gives $t^2 = t^2 + t^4$.

Philosopher. *Absolutely incredible*! Like magic, that last expression simplifies to $t^4 = 0$, so the intersection multiplicity of the parabolas' intersection at infinity is indeed $4 = 2 \cdot 2$. I can see why you're so impressed with Bézout's theorem!

The Burning Issue: Two Concentric Circles

Student. I admit there's magic, but I'm still thinking about those two concentric circles. I say they can't intersect. I'm thinking that Bézout's theorem breaks down for them. We've already agreed that these circles don't have any finite intersections.

Teacher. True, but we've just seen an example of two parabolas where the *only* intersections are at infinity, and that is what happens with two concentric circles.

Student. We have learned how to put ourselves at infinity so we can look around there, so if the circles actually meet at infinity, there must be a picture of them doing that! Can you show me one?

Teacher. Yes. To lead up to it, let me first agree with you that the whole idea of intersecting concentric circles does seem contradictory, but that's because the traditional view of circles hides some of the truth. Though no plane view tells an outright lie, it almost never tells us the whole truth, either. We can get closer to the truth by peeking into additional corners. An especially revealing choice is to look at the concentric circles in the Minkowski (x_1, iy_2)-plane. Figure 11.9 is the amazing picture of a family of concentric circles in that plane. The circles' squared radii range from positive to negative, the negative values corresponding to circles of imaginary radius.

Figure 11.9 makes one thing abundantly clear: *all the curves are approaching each other along their asymptotes*! This one picture should immediately change our thinking from "the circles don't intersect" to "they've *got* to intersect."

Student. Absolutely incredible—I guess that's what I needed to see. If I were God for just one minute, I'd make sure everyone could see in four dimensions!

Now, I still want to see actual intersection points—points I could actually color red, say.

Teacher. To arrive at such a picture, we once again do the homogenization/dehomogenization thing: the homogenization of $x^2+y^2 = r^2$ is $x^2+y^2 = r^2z^2$, and its dehomogenization obtained by putting $y = 1$ is $x^2 + 1 = r^2z^2$. For any two different values r_i, equating the right sides of $x^2 + 1 = r_i^2z^2$ yields $z = 0$. That in turn means that their common left side x^2+1 is zero—that is, $x = \pm i$. This is saying that within the viewing plane $y = 1$ (which is a complex line), any two concentric circles intersect at $x = \pm i$, $z = 0$. To see them actually intersect there, draw the parts of the circles that are in $y = 1$. Their dehomogenizations obtained by putting $y = 1$ can be written as

$$\frac{x^2}{i^2} + \frac{z^2}{\left(\frac{1}{r}\right)^2} = 1. \tag{11.2}$$

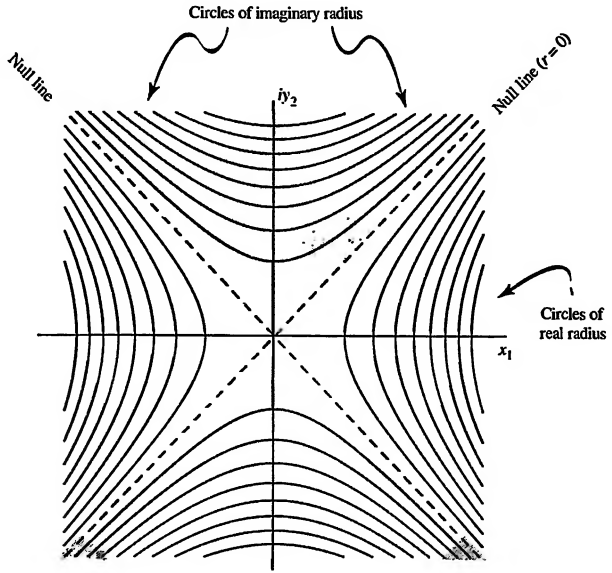


FIGURE 11.9. This is a picture in the Minkowski (x_1, iy_2)-plane of concentric circles of squared radii $r^2 = \ldots, -2, -1, 0, 1, 2, \ldots$. They look to our Euclidean eyes like "concentric hyperbolas," because the portion of a circle in the (x_1, iy_2)-plane consists of all points of constant squared separation from the origin: $r^2 = x^2 + y^2$. The dotted lines are the two null lines in the Minkowski plane that pass through the origin: the Minkowski separation from the origin to any point on these lines is zero. A line segment from the origin to any hyperbola in the unshaded region is space-like since it has positive squared separation—think of that segment as a "real radius." The hyperbolas in the shaded areas are parts of circles of imaginary radius; a line segment from the origin to any of them is time-like. The asymptotic behavior of these circles suggests that to see these circles actually intersect, we should look on this plane's line at infinity, where the circles are obviously heading.

Here is a picture of these curves, drawn in the (z_1, ix_2)-plane for choices of

$$r^2 = \ldots, -2, -1, 0, 1, 2, \ldots:$$

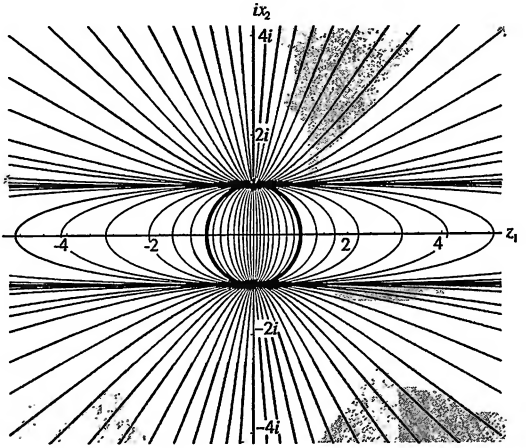


FIGURE 11.10. These are the concentric circles of Figure 11.9. The line at infinity in Figure 11.9 appears here as the vertical axis. (We've previously noted that "asymptotic" means "tangent at infinity.") This picture brings the "endpoints" of the asymptotes to center stage, showing all the circles intersecting tangentially at $(z_1, ix_2) = (0, \pm i)$. The circles intersect in two double points, making a total of four points counted with multiplicities, as Bézout's theorem claims. As in Figure 11.9, the shaded region here corresponds to $r^2 < 0$, and the unshaded region, to $r^2 > 0$. Their boundary, the two heavily-drawn horizontal lines, corresponds to $r = 0$—the null lines in Figure 11.9. Thus circles of real radius appear as ellipses, and circles of imaginary radius appear as hyperbolas. The top and bottom hyperbola branches in Figure 11.9 correspond to the top and bottom branches here, except that of course we're now seeing them from the "other end." Also, the left and right branches of Figure 11.9 correspond to the left and right ellipse halves here. Interestingly, this picture is essentially a rotated version of the "grand sweep" picture shown in Figure 8.10 on p. 136.

Student. Congratulations—what can I say? Concentric circles do actually intersect!

Interactive Applet

For an applet showing how Figures 11.9 and 11.10 are different perspectives of the same projective picture, see Chapter Applet 11 on the CD-ROM.

Philosopher. This is welcome news, but of course my hope is that there are really *four* intersection points.

Teacher. That's what Bézout's theorem says.

Philosopher. But there are actually only two points where the curves in Figure 11.10 all intersect, so it must be that any two of them intersect with multiplicity 2, making the total four.

Teacher. That must be.

Philosopher. I'm hinting at something—*can we prove this*?

Teacher. Geometrically, if an ellipse or hyperbola is tangent to another ellipse or hyperbola at a point P, then a small, appropriate translation splits up P into two nearby points.

Philosopher. Yes, but given our humble four-dimensional viewing capability, I'd feel better with an algebraic justification, as we've done before.

Teacher. I agree. For concreteness, let's concentrate on two curves in Figure 11.10—say one defined by

$$\frac{z^2}{1^2} + \frac{x^2}{i^2} = 1,$$

and the other defined by

$$\frac{z^2}{b^2} + \frac{x^2}{i^2} = 1, \quad \text{where } 0 \neq b^2 \neq 1.$$

As before, we find a parametrization of one of the curves which, when $t = 0$, gives our intersection point. Also for concreteness, of the two intersection points in Figure 11.10, let's choose $(0, i)$. Now a parametrization of the first ellipse is $z = \sin t$, $x = i\cos t$; note that when $t = 0$, this does give the point $(0, i)$. Substituting this parametrization into the equation of the second ellipse gives

$$\begin{aligned}
\frac{(\sin t)^2}{b^2} + \frac{(i\cos t)^2}{i^2} &= 1 \\
\frac{\sin^2 t}{b^2} + \cos^2 t &= 1 \\
\left(\frac{1}{b^2} - 1\right)(\sin t)^2 + (\sin t)^2 + (\cos t)^2 &= 1 \\
\left(\frac{1}{b^2} - 1\right)(\sin t)^2 &= 0 \\
(\sin t)^2 &= 0 \\
\left(t - \frac{t^3}{3!} + \frac{t^5}{5!} - \cdots\right)^2 &= 0 \\
t^2 + (\text{terms of degree} > 2) &= 0. \qquad (11.3)
\end{aligned}$$

The left side of the last line has order 2, so the intersection multiplicity is 2, which agrees with what we see geometrically.

Philosopher. I gather that the method you just used works more generally for any conics that are tangent at a point.

Teacher. Yes, whether the intersection point is finite or at infinity, and for any two ellipses, parabolas, or hyperbolas. If both curves are hyperbolas, one could use cosh and sinh in a parametrization. The power series for $\sinh t$ is

$$\sinh t = t + \frac{t^3}{3!} + \frac{t^5}{5!} + \cdots.$$

That's the same as for $\sin t$, but with all $+$'s. Similarly, the series for $\cosh t$ is a nonalternating analogue of the one for $\cos t$. These easily follow from the definitions

$$\sinh t = \frac{e^t - e^{-t}}{2} \quad \text{and} \quad \cosh t = \frac{e^t + e^{-t}}{2}.$$

> Counting intersection points isn't the only example of a "conservation law" in mathematics. Take a polyhedron drawn on a sphere. If V, E and F are the numbers of vertices, edges and faces, then $V - E + F = 2$ always. This bookkeeping holds as we remove or add vertices and edges to create other polyhedra on the sphere. An analogy from physics: an electron has charge -1, a positron has charge $+1$. When they meet and annihilate, the numbers $+1$ and -1 change. Nonetheless, the total charge stays zero. (A gamma ray is emitted, but it has no charge.) So just as V, E, and F may change, the individual charges may change, but their sum total stays the same.

The Fundamental Theorem of Algebra Revisited

Philosopher. You mentioned that Bézout's theorem is more general than the fundamental theorem of algebra. How is this?

Teacher. The fundamental theorem of algebra—we'll just call it "the fundamental theorem"—says that a degree-n polynomial f has n zeros, counted with multiplicities. Geometrically, this says that *the x-axis intersects the graph of $y = f(x)$ in n points, counted with multiplicities.* Now Bézout's theorem is a direct generalization of that, because in the statement of Bézout's theorem we can let the curve C_1 be the x-axis, which is the locus of the degree-one polynomial equation $p(x, y) = y = 0$. The other curve, C_2, is then the graph of $y = f(x)$, which is the locus of the degree-n polynomial equation $q(x, y) = y - f(x) = 0$. And of course $1 \cdot n = n$. So the fundamental theorem is Bézout's theorem for two special choices of polynomials p and q. It is instructive to make a stepwise transition from the fundamental theorem to Bézout's theorem to show just how very powerful Bézout is. For example:

- In the fundamental theorem, the x-axis can be replaced by any line, even a vertical one. One might counter that the "vertical line test" for a function says that a vertical line can intersect the graph in only one point. Bézout's theorem takes place in $\mathbb{P}^2(\mathbb{C})$, so we take C_1 and C_2 to be the topological closures in $\mathbb{P}^2(\mathbb{C})$ of the vertical line and

of the graph of $y = f(x)$. In that case, the closed vertical line intersects C_2 in a point at infinity with multiplicity $n - 1$.

Philosopher. How would you prove that?

Teacher. First, a typical vertical line in the (x, y)-plane has equation $x = a$. Second, the polynomial equation of degree n can be written as $y = a_0x^n + \cdots + a_n$. To see what happens at infinity, homogenize and dehomogenize these polynomials. Their homogenizations are

$$x = az \quad \text{and} \quad yz^{n-1} = a_0x^n + \cdots + a_{n-1}xz^{n-1} + a_nz^n,$$

and these dehomogenize at $y = 1$ to

$$x = az \quad \text{and} \quad z^{n-1} = a_0x^nz^0 + \cdots + a_{n-1}xz^{n-1} + a_nz^n. \tag{11.4}$$

The line has parametrization $x = at, z = t$. Substituting these into the second polynomial in (11.4) gives

$$t^{n-1} = a_0a^nt^n + \cdots + a_{n-1}at^n + a_nt^n. \tag{11.5}$$

Since (11.5) has order $n - 1$, the multiplicity of intersection at the "end of the y-axis" is $n - 1$, as claimed. Notice that this argument works even when $a = 0$.

- The fundamental theorem obviously applies to $y = (x - 1)(x)(x + 1)$, whose locus intersects the x-axis in the three points at $x = 1, 0, -1$. But Bézout's theorem additionally gives us greater latitude with y's exponent. For example we can make it 0, giving $1 = (x - 1)(x)(x + 1)$, which consists of three vertical lines intersecting the x-axis in three points. More interestingly, however, we can make the exponent 2 to get $y^2 = (x - 1)(x)(x + 1)$. This does nothing to the overall degree, so this curve still intersects the x-axis—as it does any line—in three points, counted with multiplicities. The locus of $y^2 = (x - 1)(x)(x + 1)$ is an example of an *elliptic* curve. The class of curves $y^2 = f(x)$ with $\deg(f) = 3$ has become somewhat famous because of its central role in A. Wiles' proof of Fermat's Last Theorem. The real locus of $y^2 = (x - 1)(x)(x + 1)$ looks like the picture on the left in Figure 11.11:

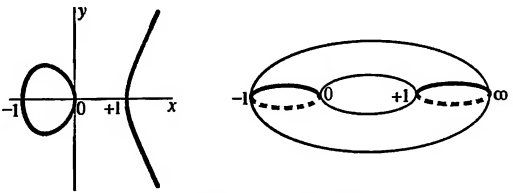


FIGURE **11.11.** The picture on the left is the real locus of

$$y^2 = (x - 1)(x)(x + 1).$$

This is an example of an elliptic curve. The entire complex curve, as a topological manifold in $\mathbb{P}^2(\mathbb{C})$, is a torus (genus 1). The real branch closes up at infinity to form a loop, illustrated by the right heavily-drawn loop on the torus.

The right portion closes up at infinity to make a second loop. As usual, such a plane drawing reveals little of the entire complex curve, which in $\mathbb{P}^2(\mathbb{C})$ turns out to be topologically a torus. The two loops are the cross-section formed when the (x_1, y_1)-plane slices through the complex curve. The loops are like the fresh-cut edges of a donut after it is cleanly sliced in half to create two C-shapes.

One can more generally replace $y = f(x)$ by $y^2 = f(x)$. By choosing f of appropriately high degree, we can create a series of loops, thus obtaining complex curves in $\mathbb{P}^2(\mathbb{C})$ of arbitrarily high genus.

- Let C_2 denote the closure in $\mathbb{P}^2(\mathbb{C})$ of $y = f(x)$'s graph. In addition to intersecting C_2 with arbitrary lines in $\mathbb{C}^2$, one can also intersect C_2 with the line at infinity. Of course the fundamental theorem has nothing to say about this, but if f is a polynomial of degree n, Bézout's theorem allows us to deduce that this intersection consists of exactly one point of multiplicity n.

Philosopher. How is that?

Teacher. In the viewing plane $y = 1$, the line at infinity has equation $z = 0$,which when parametrized by $x = t, z = 0$ and substituted into the dehomogenization

$$z^{n-1} = a_o x^n + \cdots + a_{n-1} x z^{n-1} + a_n z^n,$$

gives $0 = t^n$. So the intersection is the end of the y-axis and has multiplicity n.

- The fundamental theorem implies that the graphs of two different polynomial functions, say $y = f(x)$ and $y = g(x)$ of degrees n and m, have $\deg(f - g)$ points of intersection in $\mathbb{C}^2$. Indeed, the graphs intersect at x_o whenever $f(x_o) = g(x_o)$—that is, $(f - g)(x_o) = 0$, so that the graph of $f(x) - g(x)$ intersects the x-axis. Bézout's theorem shows that the closure in $\mathbb{P}^2(\mathbb{C})$ of the two graphs intersect at the end of the y-axis with multiplicity $nm - \deg(f - g)$.

> If $f(x)$ is a polynomial of degree n and c is a nonzero constant, then the graphs of f and $f + c$ have no finite points of intersection. Their graphs intersect in the single point at the "end" of the y-axis with multiplicity n^2.

- If one of the graphs of f, g is rotated ever so slightly, then all their intersection points become finite. Before rotation, each intersects the line at infinity at the same point—the "end" of the y-axis. After rotation, the two graphs no longer occupy the same place on the line at infinity, so there is no common point at infinity.
- The higher the degree of f, the "longer it wiggles" before calming down and moving toward a point on the line at infinity. But a transcendental function like

$$y = \sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots$$

never calms down—it wiggles all the way to infinity, acting in some ways like a "polynomial of infinite degree." (So does $y = e^x$, although we don't see it in the familiar (x_1, y_1)-plane. But in the full complex graph, it's very apparent.) Bézout's

theorem can be stretched a bit to suggest (correctly) that not only does the x-axis intersect the graph of $y = \sin x$ in an infinite number of points, but so does *any* line. If the complex line isn't "vertical" (its equation not of the form $x =$ constant) then there are infinitely many distinct points. The line $y = 1$, for example, intersects the graph at all its maxima, thus giving a series of points of multiplicity 2. In analogy with the vertically-translated parabola in Figure 5.3 [p. 71], translating this horizontal line to $y = 1 - \epsilon$ gives a series of close-together point-pairs. So does $y = 1 + \epsilon$, and just as with the parabola example, the points have separated "at right angles" and are imaginary. And what about *vertical* lines? The topological closure in $\mathbb{P}^2(\mathbb{C})$ of the complex line $x =$ constant intersects the closed complex graph of $y = \sin x$ in two points: the familiar finite one, with multiplicity 1; and a point at infinity, with infinite multiplicity—a so-called "essential singularity." This phenomenon generally holds for the complex graphs of trigonometric, exponential and logarithmic functions.

- *About the proof of Bézout's theorem*: Suppose the closure in $\mathbb{P}^2(\mathbb{C})$ of the loci in $\mathbb{C}^2$ of $p(x, y) = 0$ and $q(x, y) = 0$ both consist of n lines and m lines, respectively, and suppose there are no lines common to them. In this case, it's easy to see why Bézout's theorem holds: any one of p's lines intersects any one of q's lines in a single point, and the nm combinations—choosing one line from $p(x, y) = 0$ and one from $q(x, y) = 0$—make nm points of intersection. Now the idea of tweaking the coefficients of p or q is useful not only in separating stuck-together intersection points. By allowing appropriate coefficients to "flow to zero," it happens that one can always transform the loci to some union of lines. For example, in the ellipse $Ax^2 + Cy^2 = 1$ (A, C nonzero), letting $A \to 0$ produces $Cy^2 = 1$—two horizontal lines. Or, instead, letting $C \to 0$ gives $Ax^2 = 1$—two vertical lines. Another example: the hyperbola $xy = c (c \neq 0)$ degenerates to the union of the x- and y-axes when c becomes 0. Here's yet another: in $cy = f(x)$, decreasing the coefficient of y stretches the graph vertically. When the coefficient becomes zero, the stretch is so severe that the locus of $0 = f(x)$ has degenerated to a union of vertical lines.

Of course in a formal proof, we'd need to show that the total number of intersections remains fixed throughout this process—that the number of intersections obeys a "conservation law." (Details appear in [6, Ch. IV, §7].)

Pascal's Theorem

Philosopher. I'm eager to see how Bézout's theorem can be used to prove Pascal's Theorem.

Teacher. It's time now. We state it this way:

Pascal's theorem. Let C be an ellipse, parabola or hyperbola in $\mathbb{R}^2$. Any two opposite sides of a hexagon inscribed in C, when extended, intersect in a point of $\mathbb{P}^2(\mathbb{R})$, making for three such points. *These three points lie on a line in* $\mathbb{P}^2(\mathbb{R})$.

Actually, Pascal's theorem even continues to hold for certain degenerate conics.

Here is a picture illustrating the basic idea for an ellipse: The punch line of Pascal's Theorem is that the three points P_7, P_8, P_9 all lie on a line. This is true even if C in the picture is a parabola or hyperbola rather than an ellipse.

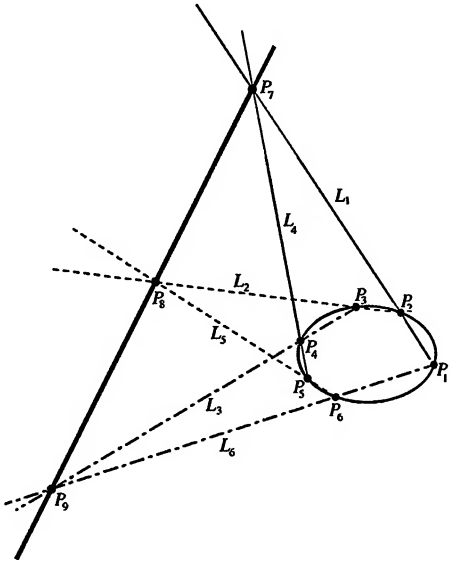


FIGURE 11.12. The six sides of a hexagon having vertices on an ellipse are extended to lines $L_1, \ldots, L_6$; the heavy line illustrates the conclusion of Pascal's theorem: the three points P_7, P_8, P_9 are collinear.

Philosopher. How could Bézout's theorem ever show this?

Teacher. You'll see! Let $P_1, \ldots, P_6$ be the six distinct points on C, and let $L_1, \ldots, L_6$ be the six distinct lines connecting pairs of successive points, as in Figure 11.12. Finally, let $C_o = L_1 \cup L_3 \cup L_5$ and $C_1 = L_2 \cup L_4 \cup L_6$. Now, as "curves," C_o and C_1 each have degree three, their equations each being the product of three degree-one polynomials. Of course, a line is defined by a degree-one polynomial.

Our proof, to my mind, has some magic to it. It's based on the fact that one can "blend" any two algebraic curves $C(p)$ and $C(q)$. We do this by blending their polynomials, and by that, I mean simply taking a linear combination of them: $\alpha p(x, y) + \beta q(x, y)$. For our purposes, we take α and β to be real. For example, if $\alpha = t$ and $\beta = 1 - t$, then as t decreases from 1 to 0, p morphs into q, and as that happens, $C(p)$ morphs into $C(q)$. Now the essence of the proof is that we can morph between C_o and C_1 in such a way that at some middle stage of this morphing, the punch line of Pascal's theorem simply falls into our laps.

Here goes. Since both C_o and C_1 consist of three lines, they intersect in $\mathbb{P}^2(\mathbb{R})$ in a total of nine points. Now in the box after Figure 4.2 [p. 60], we showed how to dehomogenize at any line of $\mathbb{P}^2(\mathbb{R})$. To simplify the argument, we choose the line at infinity so it misses

all nine points. Now not only does each of C_o and C_1 pass through all nine points, *so does each blend curve* $C(\alpha p(x, y) + \beta q(x, y))$. The reason? Any algebraic curve goes through a point P if and only if the polynomial defining it vanishes at P. So if $P \in C_o \cap C_1$, then $\alpha p(P) + \beta q(P) = \alpha \cdot 0 + \beta \cdot 0 = 0$. Now it's clear from Figure 11.12 that six of those nine points are the hexagon's vertices. But there's another observation crucial to nailing the punch line: *an arbitrary point* Q *of* $\mathbb{R}^2$ *is visited by* $C(\alpha p(x, y) + \beta q(x, y))$ *for some real* α *and* β—that is, as α and β independently sweep though all real values, the blend curve sweeps through all of $\mathbb{R}^2$. In other words, for any point Q of $\mathbb{R}^2$, there's some α and β for which $\alpha p(Q) + \beta q(Q) = 0$.

Philosopher. Specifically, *what* α and β?

Teacher. Let $\alpha = q(Q)$ and $\beta = -p(Q)$.Then $q(Q)p(Q) - p(Q)q(Q) = 0$.

Philosopher. That's pretty easy, but where does that get us?

Teacher. Well, let's choose Q to be *on* the ellipse but *off* all six vertex points. We now have *seven* points on the ellipse that are also on that cubic blend curve corresponding to this last choice of α and β. That's more than $mn = 6$, so by the second part of Bézout's theorem, the cubic blend curve must contain the ellipse as a common irreducible component.

Philosopher. I think I see it! That particular blend curve is a cubic with two components. One of its components is the ellipse, which has degree two, so the remaining part of it must have degree 1, since degrees add. So that other component is just a line. It's basically similar to what we saw with the circle-plus-x-axis a while ago.

Teacher. Exactly. The remaining three nonvertex points lie on that line, and we have our punch line.

Student. Well, I don't see what's going on. You're doing some clever things with algebra, but I'd like to be able to follow along with my eyes, seeing those two three-line curves morph between each other, changing into our conic plus a line! Maybe it works, but visually it's hard to believe how they could possibly do this.

Teacher. Your request is reasonable indeed, and I think it would be educational to plot loci corresponding to a sequence of blend polynomials, using Maple or Mathematica. The next two figures consist of a sequence of 14 pictures, showing how the morphing looks, using an ellipse as our conic. The ellipse is

$$\left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 = 1 \; (a = 3.83, b = 2.1).$$

Going counterclockwise from P_1, the inscribed hexagon has respective coordinates of approximately (0.63,2.07), (−1.06, 2.02), (−3.25, −1.11), (2.98, −1.32), (3.76, −0.41), (3.52,0.83). It turns out that $L_1 \cup L_3 \cup L_5 = C_o = C(p)$, where

$$p = (0.054x - 1.669y + 3.426)(0.208x + 6.226y + 7.603)(1.25x + 0.241y - 4.608),$$

and $L_2 \cup L_4 \cup L_6 = C_1 = C(q)$, where

$$q = (1.236x + 2.9y - 6.77)(3.13x - 2.193y + 7.73)(-0.903x + 0.776y + 3.714).$$

Figures 11.13–11.14 show the blend curves $tp + (1 - t)q = 0$ for a succession of t-values decreasing from 1 to 0.

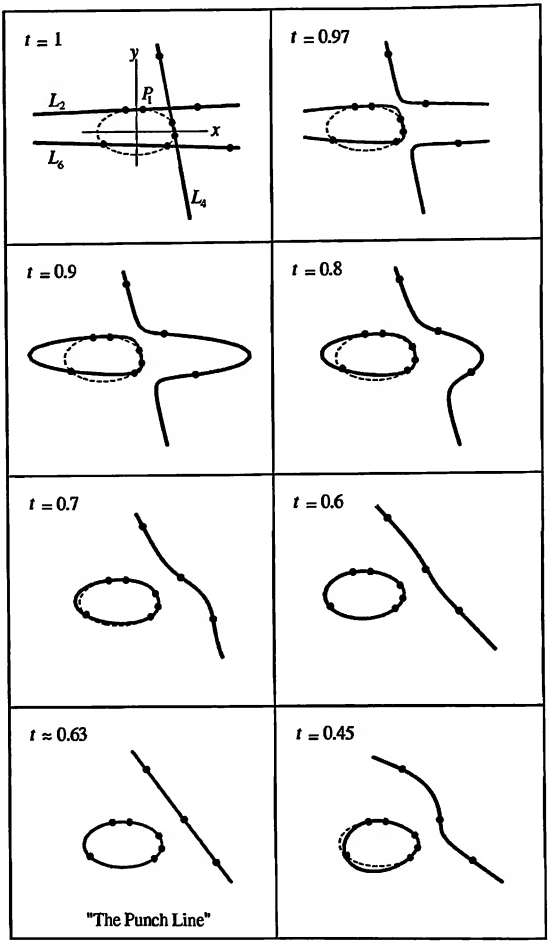


FIGURE 11.13. Top row, left shows 100% $C(p)$ and 0% $C(q)$. P_1 is the first of the six vertices, the remaining points $P_2, \ldots, P_6$, being encountered counterclockwise. $L_1 \cup L_3 \cup L_5 = C_0 = C(p)$. The "punch line" of Pascal's theorem occurs at about $t = 0.63$, when the blend curve has a seventh point in common with the original ellipse. By Bézout's theorem, that means the loop part of the blend curve is exactly the original ellipse, so the remainder of the blend curve must be a line—a degree-one curve—since the blend curve has degree three and the ellipse has degree two.

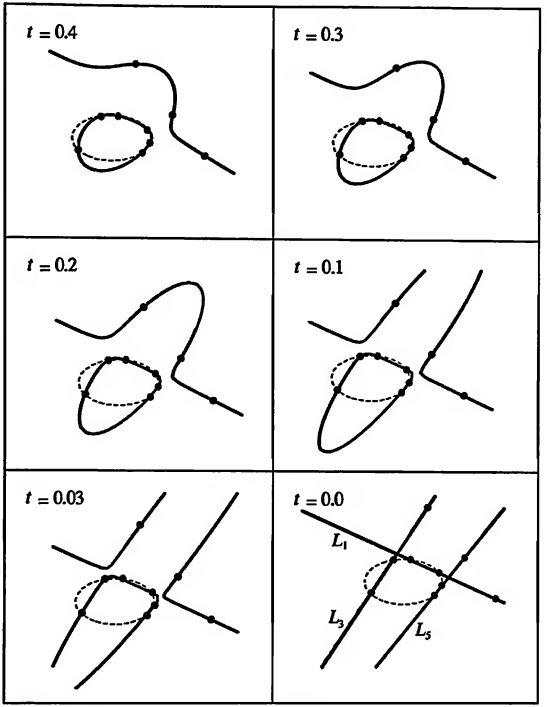


FIGURE 11.14. As $t \to 0$, the cubic approaches $C_1 = C(q) = L_2 \cup L_4 \cup L_6$.

Student. Wow! I guess seeing is believing. This stuff runs deep. I mean, it's so simple just to take linear combinations of the initial and final polynomials, but geometrically, the curves really do blend from one to the other. It *is* magic—it's amazing.

Interactive Applets

For animated versions of Figures 11.13–14, see Chapter Applets 12 and 13. For hyperbola analogues, see Chapter Applets 14 and 15 on the CD-ROM.

Some Instances of Pascal's Theorem

Teacher. Here's a sampling of seven different instances of Pascal's Theorem:

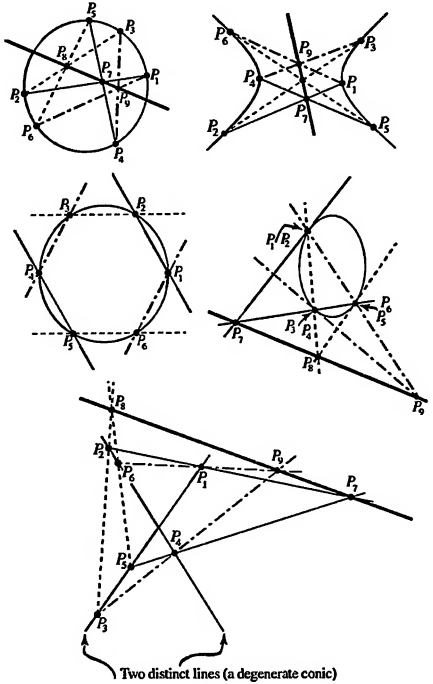


FIGURE 11.15. Top left: The hexagon need not be convex—in fact, the vertices can be in any order. Top right: Since the curve C need only have degree 2, a hyperbola works just as well. Middle left: One or all of the collinear points can be at infinity. Here the ellipse is a circle and the hexagon, regular. Opposite sides are parallel, and the three collinear points are all at infinity. Middle right: The six points of the hexagon need not all be distinct, and that means Pascal's theorem can be phrased to work for inscribed pentagons, rectangles or triangles! As two distinct points approach each other, the secant line containing them approaches the tangent line at the limit point. This picture illustrates an extreme case—three pairs of double points. Bottom: The conic can even be degenerate, this picture illustrating two crossing lines. This is actually the earliest form of Pascal's theorem—the theorem of the ancient Greek geometer Pappus. (See p. 212.) Parallel lines could also be used—they simply cross at infinity.

Interactive Applet

For an interactive version of Pascal's theorem, see Chapter Applet 16 on the CD-ROM.

Brianchon's Theorem

Teacher. It never ceases to amaze me how the one simple idea of *removing exceptions* turns from a trickle to a stream to a mighty river. I'm thinking about that famous exception in Euclidean geometry—that any two lines intersect *except* if they happen to be parallel—and how removing that exception leads to points at infinity and the projective plane.

Philosopher. Yes, and once you do that, then you can say "Any two distinct points determine a line, and any two distinct lines determine a point." The words "point" and "line" occur here in a perfectly symmetric, interchangeable way.

Teacher. And that trickle ultimately leads to the incredible realization that any fact of projective geometry can be phrased in such a way that you get a mate, or "dual," to it by interchanging "point" and "line."

Student. *Any* fact?! How can anybody be sure of that? You can't know *every* theorem. There are ones that haven't even been discovered yet! Some day someone might find an exception.

Teacher. Geometers have axiomatized projective geometry, much as Euclid axiomatized Euclidean geometry. The axioms of projective geometry can be arranged in pairs, where each axiom has a "dual," obtained by interchanging "point" and "line." The proof of any theorem involves a string of references to the axioms; the proof of the dual theorem is the same, except every reference to an axiom is replaced by a reference to the dual of that axiom.

Student. What do you mean when you say "phrased in such a way that you get a mate, or dual"?

Teacher. In mathematics and in any science, nomenclature and vocabulary typically get standardized prematurely. As just one example, if we'd known a lot more about electricity back in the days of Ben Franklin, an electron wouldn't have been assigned a charge of -1, but rather $+1$. So we have to be a bit careful to word and interpret things so the underlying symmetry is clear and transparent.

Student. Can you give me an example?

Teacher. Sure. Take a phrase like "three points on a line." The dual phrase is "three lines on a point"; by interpreting this to mean three lines all passing through one point, we have duality.

Philosopher. We just met the phrase "three points lie on a line" in the punch line of Pascal's theorem.

Teacher. Right. By the above "metamathematical" principle, every theorem of projective geometry has a dual, so...

Student. ...so Pascal's theorem has a dual?

Teacher. Yes. Pascal discovered his theorem in 1639, when he was only sixteen years old.

Remarkably, it took 167 years before its dual was discovered. A 23-year-old student at the Ecole Polytechnique in Paris, Charles-Julien Brianchon, was testing out the idea of duality, and he decided to try it out on Pascal's theorem. Brianchon's result was published in the Journal de l'Ecole Polytechnique.

Philosopher. 167 years! Amazing that it should take so long.

Teacher. No self-dual set of axioms for projective geometry was available in Pascal's time, and perhaps that is why Pascal missed it.

Philosopher. This is quite a story. I assume the punch line of that dual must be "three lines are on a point."

Teacher. Exactly. If we draw a picture illustrating the dual theorem, the main conclusion is that three lines all pass through the same point.

Student. What *is* this dual theorem?

Teacher. It says this:

Brianchon's theorem. Choose a succession of six lines tangent to an ellipse, parabola or hyperbola, the successor of the sixth line being the first line. Each pair of successive lines determines an intersection point, thus giving a succession of six points. Each pair of opposite points determines a line, and *all three of these lines pass through the same point.*

Figure 11.16 gives the idea of this theorem for an ellipse. In fact, this figure is an exact dual of Figure 11.12.

Notice how selecting six points in Figure 11.12 defines an *inscribed* hexagon, while in Figure 11.16, selecting six tangent lines defines a *circumscribed* hexagon.

Interactive Applet

For an interactive version of Brianchon's theorem, see Chapter Applet 17 on the CD-ROM.

Philosopher. I have to say the duality between Figure 11.12 and Figure 11.16 is impressive, but something bothers me. In Figure 11.12, selecting a point on the conic seems perfectly straightforward, but in Figure 11.16, there are many lines going through any particular point of the conic. What made you decide on the *tangent* line?

Teacher. Your question goes directly to a deeper point/line property. To put it succinctly, if a conic can be looked at as a collection of its *points*, then if duality is so pervasive, we should be able to regard the conic as a collection of its *lines*.

Student. This sounds crazy. What do you mean by "*its lines*"?

Teacher. Well, here's some intuition about this. Look at an ellipse, say, as composed of a succession of one million very closely-spaced points. Any two adjacent points determine the line containing them, and that line is a secant line looking very much like a tangent line to the ellipse.

Philosopher. So that's where the tangent lines are coming from?

Teacher. That's the idea. Now if we construct and plot all one million lines corresponding to the one million adjacent point-pairs, you'd see the ellipse as the "envelope" of all these

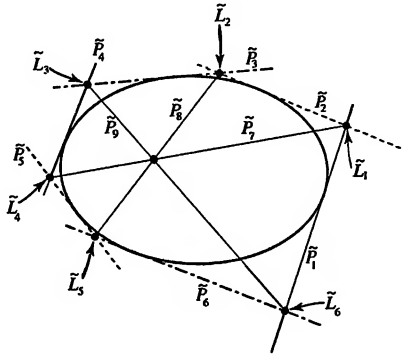


FIGURE 11.16. This picture was constructed by applying a point/line "dualizing operation"—which we denote by "$\sim$"—to Figure 11.12. The construction begins by replacing each point P_i in Figure 11.12 with the line tangent to the conic at P_i, denoted by $\tilde{P}_i$. Next, in analogy to Figure 11.12 (where two successive points determine a line), two successive lines here determine a point. Finally, in analogy to Figure 11.12 (where the three pairs of opposite hexagon sides determine three points), the three pairs of opposite hexagon vertices here determine three lines. The heavy dot symbolizes the conclusion of Brianchon's theorem: these three lines are concurrent. This dot stands in analogy to the heavy line in Figure 11.12 highlighting the conclusion of Pascal's theorem that the three points are collinear.

lines. Plotted on paper, it would look like a white "eye" popping out at you from the middle of a black sea. And this process can be dualized: in an ellipse consisting of a succession of one million lines, any two lines adjacent in this succession determine a point. Throwing away all the lines and just staring at these intersection points gives a figure looking almost exactly like the original ellipse! So that is what's going on. One can of course make things more rigorous by using more and more points, ever closer together, and appropriately taking the limit of this process.

Philosopher. This seems to be saying that a conic is "self-dual"—it can be regarded equally well as a collection of points or of (tangent) lines.

Teacher. Yes. In going from either of these theorems to the other, we're just toggling between two ways of looking at a conic.

Just a minute ago, we were struck by the long time span between the two theorems. But in the degenerate case of two lines, Pascal's theorem was noticed—and proved—much earlier.

Student. How much earlier?

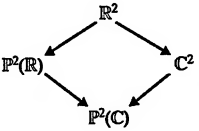
Teacher. Some thirteen centuries! It is a famous theorem due to Pappus of Alexandria—the last of the great Greek geometers. Historians generally estimate that he flourished around 290–350 C.E. This is his theorem:

Theorem of Pappus. Take two lines, and three points on each line. Alternately connect them to make a hexagon. *The three pairs of opposite sides of this hexagon intersect in three points lying on a straight line.*

Interestingly, Pappus apparently took quite an interest in hexagons. He wrote admiringly about how bees wisely choose a regular hexagonal shape for honeycombs, noting that this shape uses less material than a honeycomb based on squares or equilateral triangles.

Chapter Summary

In the diagram



the completion $\mathbb{R}^2 \to \mathbb{P}^2(\mathbb{R})$ removes exceptions arising when points escape to infinity, and clears the way to a beautiful and powerful duality between points and lines. On the other hand, $\mathbb{R}^2 \to \mathbb{C}^2$ removes exceptions arising when points escape into the complex world, and makes possible the beautiful and powerful fundamental theorem of algebra. The final transition to $\mathbb{P}^2(\mathbb{C})$ banishes *both* sorts of exceptions, and makes possible the beautiful and powerful Bézout theorem. As in the fundamental theorem of algebra, one must count intersection points "with multiplicities," and in this chapter we look at how to do this in two ways. In the geometric way, we suitably perturb one or both curves to separate "stuck together" points so they can be counted. In the algebraic way, we substitute into the equation of one curve a parametrization of the other curve, this parametrization being centered at the intersection point P; the resulting order in the parametrization variable is the intersection multiplicity at P. The chapter is completed with some basic applications of Bézout's theorem: identifying all four intersection points of two concentric circles; Pascal's famous "hexagon theorem"; and its dual, Brianchon's theorem.

Exercises

1. Figure 11.8 on p. 195 shows how rotating either of the cubics $y = x^3$ and $y = 2x^3$ separates their triple intersection point; however, translating either one vertically yields but one real intersection point. Have Maple or Mathematica find all three intersection points under a small vertical translation. Do the same for a small horizontal translation.
2. For each of the following pairs of curves, find, using Maple or Mathematica, all intersection points guaranteed by Bézout's theorem.
   (a) The two circles $(x-1)^2 + y^2 = 4$ and $(x+1)^2 + y^2 = 4$;
   (b) $y = x + x^3$ and the line $y = mx$ for $m = 0$; $m = 1$; $m = \infty$;
   (c) $x^2 + y^2 = 1$ and $y = x^3$;
   (d) The cuspidal cubics $y^2 = x^3$ and $y^3 = x^2$.

3. (a) Use Maple or Mathematica to find the four points in which the circle

$$(x-\epsilon)^2+y^2=1$$

intersects the circle $x^2+y^2=4$, where ϵ is a small real number.

(b) Show, by running (a) for a series of ever-smaller ϵ's, that your four intersection points get ever-closer to the ones found in this chapter for the concentric circles $x^2+y^2=1$ and $x^2+y^2=4$.

4. We have shown that if $p(x)$ is a polynomial of degree n, then the closure in $\mathbb{P}^2(\mathbb{C})$ of the complex graph of $y=p(x)$ intersects the line at infinity in a single point of multiplicity $n-1$. That means that as $c\to\infty$, the complex line $y=c$ intersects that graph in $n-1$ points that in some sense converge to that single multiple point at infinity. Position yourself at infinity, watch the spectacle, and describe what you see.

5. (a) Use Maple or Mathematica to sketch a series of blend curves starting with the graph of $y=\cos x$ and ending with the graph of $y=\sin x$.

(b) Sketch a series of blend curves starting with the graph of $y=\cos x$ and ending with the circle $x^2+y^2=1$.

6. Figure 11.15 (p. 208) includes a picture illustrating Pascal's theorem when the conic is a hyperbola. Use Maple or Mathematica to create a series of sketches based on a hyperbola, complementing the series of ellipse sketches in Figures 11.13–11.14.

7. Modify the language of our proof of Pascal's theorem so it directly proves the theorem of Pappus.

8. The Three Tangents Theorem: Suppose a triangle with vertices a, b, c is circumscribed about an ellipse, touching the ellipse at points A, B, C on the sides opposite a, b, c. Prove that the lines Aa, Bb, Cc are concurrent.

9. State and illustrate the "inscribed triangle" dual to the Three Tangents Theorem.

10. (a) By eliminating t, show that the "Lissajous figure" parametrized by $x=\cos 2t$; $y=\sin t$ is the locus of a polynomial equation. What is its degree?

(b) Find all the intersection points guaranteed by Bézout's theorem for this Lissajous figure and the line $y=0$;

(c) Same as (b), for the line $y=x$;

(d) Same as (b), for the line $x=0$.

11. To determine the multiplicity of an intersection point, we've used perturbations as well as substituting into the equation of one curve a parametrization of the other curve. A third approach based on linking numbers is suggested in the box on p. 193. A command in Maple's algebraic curves library allows one to explore this idea—the "plot_knot" command draws the intertwined curves mentioned in the box, and in simple cases one can visually check that the winding number is indeed the intersection multiplicity obtained using the perturbation or substitution method. We explore this in the following examples.

(a) Use the substitution method to determine that the cuspidal curve $y^2=x^3$ and the y-axis intersect at the origin with multiplicity 2.

(b) Access Maple's algebraic curves library by entering

with(algcurves);

If you write the cusp as $y^2 - x^3 = 0$ and the y-axis as $x = 0$, then their union is the locus of $x(y^2 - x^3) = 0$. Maple can plot the intersection of each curve with the 3-sphere centered at the origin. Here's the command, which should be entered on one line:

plot_knot(x*(y^2-x^3),x,y,epsilon=0.5,radius=0.07,

color=[green,red],orientation=[0,0]);

The radius of the 3-sphere in $\mathbb{R}^4$ is specified by "epsilon=0.5" and the radius of the string or tube used to represent the curves is given by "radius=0.07." One loop is colored green, the other red. Every full winding of, say, the green curve around the red results in the green going over red once, and under red once. The picture should show a total of two "overs" and two "unders" for each curve relative to the other. Although the initial orientation has been chosen in this command line, one can select the picture and drag to rotate the pair of linked loops, thus obtaining a better sense of their three-dimensional arrangement.

(c) Verify algebraically that $y^2 - x^3 = 0$ intersects the x-axis with multiplicity 3.

(d) Enter on one line:

plot_knot(y*(y^2-x^3),x,y,epsilon=1.0,radius=0.07,

color=[green,red],orientation=[0,0]);

See that their mutual winding number is 3.

12

The Big View

in which families of 3-D slices show us, in different ways, the big picture of complex conics

What Does a Complex Circle Look Like?

Student. That experience of seeing concentric circles intersect makes me realize that we don't see much of *any* conic! Didn't we once say that a complex circle is actually *two* dimensional? The most we've seen are parts in 3-D slices, and each of those looks like a wire model—one-dimensional.

Teacher. Yes, we'll see that any complex circle is in fact two-dimensional.

Philosopher. Actually, so far we've seen only the "tips of a mountain range." We've constructed a few 3-D slices, allowing us to see enough to draw the conclusions we needed at the time. The whole mountain range is the entire complex conic sitting in $\mathbb{P}^2(\mathbb{C})$. That complex conic has real dimension 2, and it lives in $\mathbb{P}^2(\mathbb{C})$, which has real dimension 4. I don't really know *what* the whole mountain range looks like!

In addition, the 3-D slices we've chosen are all very special. For example, take our unit circle $x^2 + y^2 = 1$, and suppose we walk from the origin along the x_1-axis. When we encounter the circle, we can begin climbing up a hyperbola branch in (x_1, iy_2)-space. Our choice of the x_1-axis leads to (x_1, y_1, iy_2)-space, the slice showing both the real circle and the attached hyperbola. Similarly, choosing the y_1-axis leads to (x_1, ix_2, y_1)-space as the most informative 3-D slice, and there's again an attaching hyperbola. But these choices of travel are along *axes*. Couldn't we just as well start at the origin and walk along *any* line?

Teacher. You have this incredible way of pointing out how arbitrary we've been! You're absolutely right, and I have to confess that I don't know *what* we'd see if we walked from the origin along a different line.

Philosopher. It seems to me that, in addition, there should be a smooth transition between these two 3-D pictures—the one in (x_1, y_1, iy_2)-space and one in (x_1, ix_2, y_1)-space.

Student. The two pictures look basically the same to me!

Philosopher. Yes, but notice that the vertical axis in one is iy_2, while in the other it's ix_2. How do you morph from one to the other?

Teacher. I don't know, but I do think it's time to apply a bit of mathematical wisdom: *When algebra runs dry, use geometry; when geometry runs dry, use algebra.* In our case, geometry has run pretty dry—it seems that none of us can see the "whole elephant," so I suggest we start by using the power of algebra. I'm thinking of using the parametrization $x = \cos t$, $y = \sin t$ of the unit circle, because t can be regarded as an angle, and fixing it determines a ray from the origin.

> "Trying the opposite end" often leads to progress in mathematics. *When algebra runs dry, use geometry; when geometry runs dry, use algebra* is just one example. Here's another: *If you keep running up against a wall in trying to prove something in mathematics, try constructing a counterexample.* If your conjectured theorem is wrong, you may, of course, find a counterexample. If it's right, then constant failure in attempts to construct a counterexample can itself provide helpful information. Gödel and Abel both tried the opposite end, succeeding in establishing unexpected negative results after years of attempts by brilliant mathematicians to prove the positive.

Philosopher. Looked at that way, the only slices we've used correspond to t being a multiple of $\frac{\pi}{2}$.

Teacher. Yes. But we've seen before that the identity $\cos^2 t + \sin^2 t \equiv 1$ remains true for t complex, and it turns out that using $t = t_1 + it_2$ provides a parametrization of the entire complex circle. We can write

$$\begin{aligned} x &= \cos(t_1 + it_2) = \cos t_1 \cdot \cos it_2 - \sin t_1 \cdot \sin it_2; \\ y &= \sin(t_1 + it_2) = \sin t_1 \cdot \cos it_2 + \cos t_1 \cdot \sin it_2. \end{aligned} \tag{12.1}$$

Student. I don't know what $\cos it_2$ and $\sin it_2$ are.

Teacher. Well, we can easily put them into a more familiar form. Using Euler's identity, write

$$\begin{aligned} e^{i\theta} &= \cos\theta + i\sin\theta, \\ e^{-i\theta} &= \cos\theta - i\sin\theta. \end{aligned}$$

By adding and subtracting these equations, we get cosine and sine:

$$\begin{aligned} \frac{e^{i\theta} + e^{-i\theta}}{2} &= \cos\theta; \\ \frac{e^{i\theta} - e^{-i\theta}}{2} &= i\sin\theta. \end{aligned}$$

Philosopher. Those left-hand sides look like cosh and sinh, with arguments $i\theta$, don't they?

Euler (1707–1783) discovered his famous identity $e^{i\theta} = \cos\theta + i\sin\theta$ in 1748. It led him to

$$e^{i\pi} + 1 = 0,$$

often called "the most remarkable formula in mathematics." It links in one equation five of the most important constants in mathematics: 0, 1, i, π, e.

Teacher. They do. If we substitute it for θ, we get

$$\cos it = \frac{e^{-t} + e^{t}}{2} \quad \text{and} \quad i\sin it = \frac{e^{-t} - e^{t}}{2}.$$

We can therefore rewrite these last two equations as

$$\cos it = \cosh t; \quad \sin it = i\sinh t. \tag{12.2}$$

So our circle's parametrization (12.1) becomes

$$\begin{aligned} x &= \cos(t_1 + it_2) = \cos t_1 \cosh t_2 - i\sin t_1 \sinh t_2; \\ y &= \sin(t_1 + it_2) = \sin t_1 \cosh t_2 + i\cos t_1 \sinh t_2. \end{aligned} \tag{12.3}$$

We can write this in component form in $\mathbb{R}^4$:

$$(x_1, x_2, y_1, y_2) = (\cos t_1 \cosh t_2, -\sin t_1 \sinh t_2, \sin t_1 \cosh t_2, \cos t_1 \sinh t_2). \tag{12.4}$$

Philosopher. Am I right in saying that as t_1 and t_2 vary independently, (12.4) paints a picture of the complex unit circle in $\mathbb{R}^4$?

Teacher. Yes. The parameter-pair (t_1, t_2) defines a plane, and (12.4) maps that plane into our two-dimensional complex circle. Notice that the original circle corresponds to real t. That is, setting $t_2 = 0$ means $t = t_1$ and then, because $\cosh 0 = 1$ and $\sinh 0 = 0$, (12.4) becomes

$$(\cos t_1, 0, \sin t_1, 0),$$

the familiar parametrization of the real circle. On the other hand, setting $t_1 = 0$ means that $t = it_2$, and (12.4) becomes

$$(\cosh t_2, 0, 0, \sinh t_2). \tag{12.5}$$

This is none other than a parametrization of the right hyperbola branch attached to the real unit circle. Since the functions appear as the first and fourth coordinates, the branch lives in the (x_1, iy_2)-plane, as it should.

Student. And the left branch?

Teacher. Set $t_1 = \pi$. Since $\cos \pi = -1$, equation (12.4) becomes

$$(-\cosh t_2, 0, 0, -\sinh t_2). \tag{12.6}$$

This sweeps out the left branch as t_2 runs through the reals.

Philosopher. It seems that fixing t_1 determines a ray from the origin in the ordinary real (x_1, y_1)-plane, and if we then let t_2 vary, we get an attaching hyperbola branch. I wonder if that works for the vertical rays? Does (12.4) give us the two hyperbola branches in (ix_2, y_1)-space?

Teacher. Yes, it does. Setting $t_1 = \pm\frac{\pi}{2}$ means $\cos t_1 = 0$ and $\sin t_1 = \pm 1$, and then (12.4) reduces to

$$(0, \mp \sinh t_2, \pm \cosh t_2, 0), \tag{12.7}$$

which gives exactly those branches in (ix_2, y_1)-space.

Philosopher. I'm impressed! (12.4) seems really powerful, and I am eager to see if it can shed light on those other directions. That is, if we choose *any* value for t_1, will it reveal what happens as we continue traveling along the ray through the circle's real point $(\cos t_1, 0, \sin t_1, 0)$?

Teacher. We may rewrite (12.4) as a linear combination of two vectors, like this:

$$\cosh t_2(\cos t_1, 0, \sin t_1, 0) + \sinh t_2(0, -\sin t_1, 0, \cos t_1). \tag{12.8}$$

The vectors $(\cos t_1, 0, \sin t_1, 0)$ and $(0, -\sin t_1, 0, \cos t_1)$ are orthonormal in $\mathbb{R}^4$; as t_1 varies, each of them traces out an ordinary circle in $\mathbb{R}^4$ In $\mathbb{C}^2$ it's easy to see what the first vector does. It traces out our ordinary circle in the real (x_1, y_1)-plane. The second vector also traces out a circle, but it's in the pure imaginary (ix_2, iy_2)-plane. In the first plane, all separations-squared are nonnegative; we have a Euclidean circle in a Euclidean plane. In the second plane, all separations-squared are nonpositive, and we may say that we have an "anti-Euclidean" circle in an anti-Euclidean plane.

Student. Where are you going with this?

Teacher. Suppose that as t_1 varies, you suddenly freeze it at t_o. We then have two corresponding frozen vectors

$$\mathbf{v} = (\cos t_o, 0, \sin t_o, 0) \quad \text{and} \quad \mathbf{w} = (0, -\sin t_o, 0, \cos t_o).$$

Here $\mathbf{v}$ is in Euclidean space, $\mathbf{w}$ is in anti-Euclidean space, they're orthogonal, and together they $\mathbb{R}$-span a Minkowski plane in $\mathbb{C}^2$. Writing (12.8) as

$$\cosh t_2\, \mathbf{v} + \sinh t_2\, \mathbf{w} \tag{12.9}$$

shows that as t_2 varies, we get a hyperbola branch in that Minkowski space. When $t_2 = 0$, it touches our real circle at the point $\mathbf{v} = (\cos t_o, 0, \sin t_o, 0)$.

Philosopher. Does this generalize what we've seen before?

Teacher. Yes, it does, because if t_1 is frozen at 0 (that is, $t_o = 0$), then

$$\mathbf{v} = (1, 0, 0, 0) \quad \text{and} \quad \mathbf{w} = (0, 0, 0, 1),$$

and we get the branch in (12.5). When $t_o = \pi$, we have

$$\mathbf{v} = (-1, 0, 0, 0) \quad \text{and} \quad \mathbf{w} = (0, 0, 0, -1),$$

giving the branch in (12.6). Set $t_o = \pm\frac{\pi}{2}$ and we have

$$\mathbf{v} = (0, 0, \pm 1, 0), \mathbf{w} = (0, 0, \mp 1, 0),$$

giving the branches in (12.7).

Philosopher. I get the idea that in all our previous 3-D slices, v defines the real axis, and w defines the imaginary one. The hyperbola always attaches to the real circle, but itself lives in a Minkowski plane, which is what v and w always generate. It seems that (12.9) generalizes this idea.

Teacher. That's right. And you said earlier that there ought to be a smooth transition between the two 3-D pictures, the one in (x_1, y_1, iy_2)-space and one in (x_1, ix_2, y_1)-space. Well, v and w show how. As v rotates around the real circle from the x_1-axis to the y_1-axis, w rotates around the anti-Euclidean circle from the iy_2-axis to the ix_2-axis, $\frac{\pi}{2}$ out of phase with v. At each stage a hyperbola attaches to the circle, and if we extend the ordinary real (x_1, y_1)-plane to a 3-space that includes w, we can see the attached hyperbola. Here is a picture illustrating this:

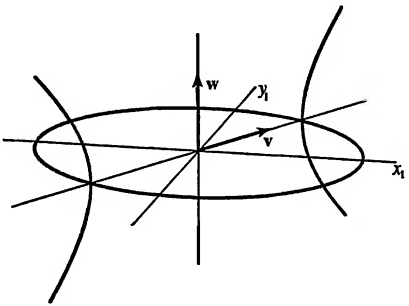


FIGURE 12.1. As the vector v rotates around an ordinary real circle in the Euclidean (x_1, y_1)-plane, a "dual" vector w rotates around another circle in the anti-Euclidean (ix_2, iy_2)-plane. The pair v and w, by determining a special family of 3-D slices of $\mathbb{R}^4$, permits us to see what inhabitants of $\mathbb{R}^4$ have known all along: to every point of the real circle there is attached a hyperbola branch. Therefore what we've seen up to now—v lying on the x_1 or y_1-axis—is just part of a larger picture. The plane containing v and w is always Minkowski.

Philosopher. So at *any* point of our real circle, there's a hyperbola branch attached to it! To me, the big question seems to be, *where is it*? I'm thinking that it's in the plane containing v and w.

Teacher. I agree. Notice that since **v** and **w** are orthonormal, (12.9) shows that all the branches have the same shape. That is, what we see in any of the (**v**, **w**)-planes, with our Euclidean eyes, always looks the same. Also, notice that as **v** moves, so does **w**, and therefore so does the plane containing them. In fact, the 3-space containing the circle and attached hyperbola rotates about the real (x_1, y_1)-plane.

Student. You've totally lost me! What on earth do you mean by a 3-space rotating about an ordinary real plane?

Teacher. Let me try to explain with an analogy. The basic idea underlying what I'm about to say is the fancy word "cylindrization," but the idea is quite simple. Take the ordinary Euclidean (x, y)-plane: we can cylindrize any subset S in it by passing through each point of S a line in the vertical direction. So a point becomes a vertical line, a circle becomes a vertical cylinder, a line segment becomes a vertical strip, and so on.

Philosopher. Just as my far-out high school teacher did, once. [See p. 13.]

Student. So a square becomes a vertical pipe with square cross section, a triangle becomes a pipe with a triangular cross section, and a line becomes a plane parallel to the z-axis?

Teacher. You've got it. If that line in the (x, y)-plane goes through the origin, the plane contains the z-axis. And as the line turns about the origin, the plane turns about the z-axis! All these cylindrizations do is jack up the dimension by 1—that's it. Any activity in the plane, be it translation, turning, or whatever, extends vertically. And its shadow on the z-axis is just the original plane scenario.

Now there's nothing to stop us from cylindrizing again, into the fourth dimension. Maybe we don't see it, but on the other hand, that's really the *only* difficulty. Everything else about it is straightforward. In this fancier cylindrization, a point becomes a plane, and a line becomes a 3-space. A line in the (x, y)-plane rotating around a point there becomes a 3-space rotating around a plane. By putting the sun at a distant point of the t-axis, the shadow of all this is in (x, y, z)-space, where we see a plane rotating around a line, weather vane-style. Casting a shadow of this activity in (x, y, z)-space into the original Euclidean plane just gives what we started with—a line spinning around the origin!

A Topological Model

Philosopher. So we have a hyperbola branch attached to each point of the real circle. But how do all these branches fit together?

Teacher. Notice that all the branches race off to infinity.

Philosopher. Sure.

Teacher. In $\mathbb{P}^2(\mathbb{C})$, each branch meets the complex line at infinity in two points.

Student. One point for each of the two "ends" of the branch?

Teacher. Yes.

Student. So as we move around the circle, encountering an attached branch at each point, we get two corresponding curves in the complex line at infinity?

Teacher. Not at all. We get two *points*, not two curves.

Student. Give me a break! All those branches are heading off in different directions, even going into different 3-spaces. How could they converge to just two points at infinity?

Teacher. Take two people at the south pole. They could start off in completely different directions, but if each travels along a line of longitude, their paths will eventually meet at the north pole, and that's a point.

Philosopher. Aren't you sort of pulling a rabbit out of a hat, here?

Teacher. It all comes from Bézout's theorem.

Student. How, exactly?

Teacher. Remember, we're talking about one big object, a complex circle, and those hyperbola branches are just one-dimensional parts of it. By Bézout's theorem the circle, being defined by a polynomial of degree 2, intersects any complex line, including the one at infinity, in two points. There are lots of different paths going out to infinity, just as there are lots of ways of going from the south pole to the north pole. How we decide to go to infinity is our business, but Bézout's theorem says that if it's infinity we're headed for, then we'll land at exactly one of two points—those very two points in which our infamous concentric circles meet.

Student. *Those*! I just knew we'd meet them again, somehow.

Philosopher. I'm still trying to see just how all the hyperbolas fit together.

Teacher. We can get a topological model of the entire complex circle in $\mathbb{P}^2(\mathbb{C})$ this way: we know one branch attaches to each point of a real circle, and that exactly two points are common to all these branches. The lines of longitude I mentioned a moment ago seem to provide an answer: all of them meet at two points, which we can call the "north pole" and "south pole." Figure 12.2 gives the idea:

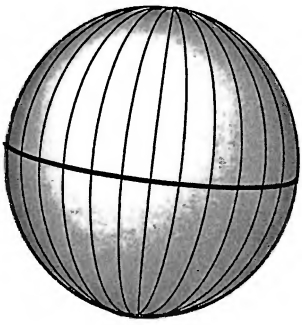


FIGURE 12.2. All those hyperbola branches attached to a real circle topologically fit together to form this picture. The "equator" is the real circle, and the lines of longitude are the attached hyperbola branches. All these branches pass through the same two points of the complex line at infinity of $\mathbb{P}^2(\mathbb{C})$. These points appear here as the north and south poles of a sphere.

To make a model of the complex circle fit into $\mathbb{R}^3$, one has to do some massaging, but it turns out that the sphere is topologically faithful. Each branch is unbounded in $\mathbb{C}^2$; we've shrunk them so they're bounded. Any branch lies in a plane both before and after massaging, but in our model, all the planes are in $\mathbb{R}^3$and any two of them intersect in a line—the line containing the north and south poles. In $\mathbb{P}^2(\mathbb{C})$ or in $\mathbb{C}^2$ the planes have more freedom to move around, and they turn in such a way that any two of these hyperbola-containing planes intersect in only the origin.

Philosopher. How would you prove that?

Teacher. Take two different instants of time, giving

$$\mathbf{v} = (\cos t_o, 0, \sin t_o, 0),$$
$$\mathbf{v}' = (\cos t_o', 0, \sin t_o', 0),$$
$$\mathbf{w} = (0, -\sin t_o, 0, \cos t_o),$$
$$\mathbf{w}' = (0, -\sin t_o', 0, \cos t_o').$$

Since the v's all terminate on the real unit circle, v and v′ are linearly independent as long as $t_0 - t_o'$ is not an integral multiple of π. The same holds for w and w′. And the position of the 0-components of these 4-vectors implies that any nonzero linear combination of the v's is different from any nonzero linear combination of the w's. From this it's easy to conclude that all four vectors are linearly independent. Thus the 2-space containing any two of the four vectors intersects the 2-space containing the other two vectors in only the origin.

The real portion of a complex circle is misleading, because it shows only the equator of the sphere in Figure 12.2. A curve like $x^{100} + y^{100} = 1$ provides a good example of how *tremendously* misleading the view in the real plane can be. In $\mathbb{R}^2$, the locus of this polynomial equation looks just like a square, though upon close inspection, one sees a slight rounding at the corners. In truth, however, the locus of

$$x^{100} + y^{100} = 1 \quad \text{in} \quad \mathbb{P}^2(\mathbb{C})$$

is a smooth surface—a donut with not just one, but 4851 holes in it! In $\mathbb{R}^2$, the curve $x^{1000} + y^{1000} = 1$ simply looks, under scrutiny, like a square with sharper corners. In $\mathbb{P}^2(\mathbb{C})$, however, it looks like a donut with 498,501 holes. The number of holes is called the genus, denoted by g, and one can topologically visualize the surface as a sphere with g handles.

A "Conjugate" Model

Philosopher. Figure 12.2 is beautiful—a nice picture of all those hyperbola branches meeting at the sphere's poles. But I've been thinking...we selected each branch by fixing t_1. As t_2 varies, aren't we then tracing out a meridian—a line of longitude on the sphere? It

seems to me that under the parametrization (12.1) [p. 216], the vertical lines in the complex t-plane map onto those lines of longitude. Figure 12.3 expresses what I'm saying:

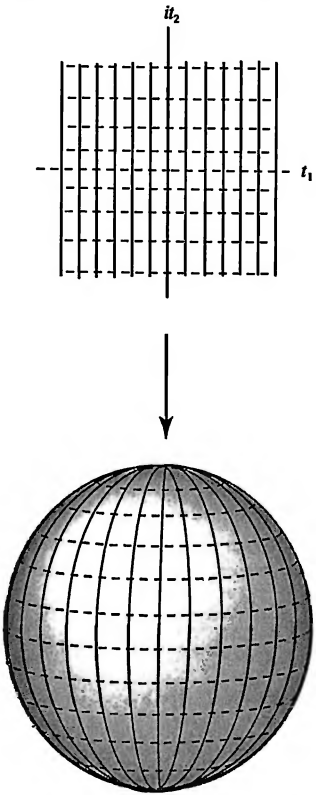


FIGURE 12.3. Varying just one of t_1 and t_2 in $t = t_1 + it_2$ produces "coordinate lines" in the parameter space $\mathbb{C}$. These lines map under the parametrization (12.1) to parallels of latitude and lines of longitude on a sphere.

As you move from one vertical line to another, you go from one line of longitude to another. And if you increase t_1 by 2π, you're back to the same line of longitude.

Teacher. I agree. After all, both sin and cos have period 2π.

Philosopher. So actually, a vertical strip of width 2π maps to the sphere, and successive strips of width 2π just keep wrapping around the sphere.

Teacher. That's right.

Philosopher. Then I begin to wonder: if a real circle has a hyperbola branch attached to it at each point, and if these all fit together as lines of longitude to make a sphere, then is there a "dual" picture for a real *hyperbola*? Might there be a circle attached to each of its points, and do they fit together as circles of latitude to make a sphere, too?

Teacher. What a natural question! I'll bet there is such a dual picture, and I'm guessing we can use the same approach as before.

Hyperbolic functions have addition formulas, too, and here they are:

$$\cosh(u+v) = \cosh u \cdot \cosh v + \sinh u \cdot \sinh v;$$

$$\sinh(u+v) = \sinh u \cdot \cosh v + \cosh u \cdot \sinh v.$$

To derive them, convert the right-hand sides to their exponential definitions, expand and collect terms; what remains amounts to the exponential definitions of the left-hand sides.

Choosing $u = t_1$ and $v = it_2$ leads to

$$\cosh(t_1 + it_2) = \cosh t_1 \cdot \cos t_2 + i \sinh t_1 \cdot \sin t_2;$$

$$\sinh(t_1 + it_2) = \sinh t_1 \cdot \cos t_2 + i \cosh t_1 \cdot \sin t_2.$$

Before, the circle's parametrization in (12.1) [p. 216] led to (12.4), so now let's use the following hyperbola-branch parametrization, which should lead to an analogue of (12.4):

$$\begin{aligned} x &= \cosh(t_1 + it_2) = \cosh t_1 \cdot \cos t_2 + i \sinh t_1 \cdot \sin t_2; \\ y &= \sinh(t_1 + it_2) = \sinh t_1 \cdot \cos t_2 + i \cosh t_1 \cdot \sin t_2. \end{aligned} \tag{12.10}$$

In component form, our analogue to (12.4) is

$$(x_1, x_2, y_1, y_2) = (\cosh t_1 \cdot \cos t_2, \sinh t_1 \cdot \sin t_2, \sinh t_1 \cdot \cos t_2, \cosh t_1 \cdot \sin t_2).$$

We can rewrite this so the vectors depend only on t_1:

$$\cos t_2(\cosh t_1, 0, \sinh t_1, 0) + \sin t_2(0, \sinh t_1, 0, \cosh t_1); \tag{12.11}$$

this is an analogue to (12.8).

Now if we freeze t_1 at some T_o, we get two vectors

$$\mathbf{V} = (\cosh T_o, 0, \sinh T_o, 0) \quad \text{and} \quad \mathbf{W} = (0, \sinh T_o, 0, \cosh T_o).$$

In analogy to (12.9) we can write (12.11) as

$$\cos t_2\,\mathbf{V} + \sin t_2\,\mathbf{W}. \tag{12.12}$$

Philosopher. I notice that just as **v** and **w** are orthogonal, so are **V** and **W**. But they don't seem to have unit length like **v** and **w**.

Teacher. That's right. In $\mathbb{R}^4$, **V** and **W** both have squared distances $\cosh^2 T_o + \sinh^2 T_o$, and that turns out to be $\cosh 2T_o$. So although **V** and **W** are equally long in $\mathbb{R}^4$, their lengths increase without bound as $|T_o| \to \infty$.

Philosopher. So what do we see, geometrically?

Teacher. In $\mathbb{R}^4$, pass a plane through **V** and **W**. Since **V** and **W** are orthogonal and have the same length, in that plane they act much like orthonormal vectors in the (x, y)-plane, in that $\cos t_2\mathbf{V} + \sin t_2\,\mathbf{W}$ sweeps out a circle in the plane as t_2 advances by 2π. Here the circle's radius is $\sqrt{\cosh^2 T_o + \sinh^2 T_o}$.

Student. That's a space-circle in $\mathbb{R}^4$?

Teacher. Yes.

Philosopher. It seems that the hyperbola branch traced out as t_1 varies now plays the same role as the real circle traced out by t_1 did earlier. That is, just as Figure 12.1 suggests that there is attached a hyperbola branch to each point of the real circle, Figure 12.4 analogously suggests that a circle is attached to each point of the real hyperbola branch:

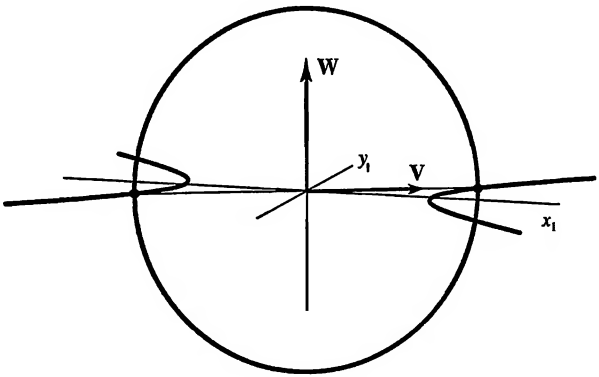


FIGURE 12.4. This is the companion to Figure 12.1. Here the vector **V** swings around the real hyperbola regarded as a loop, and each **V** in the Euclidean (x_1, y_1)-plane determines a dual vector **W** in the anti-Euclidean (ix_2, iy_2)-plane. The pair **V**, **W**, by determining just the right family of 3-D slices of $\mathbb{R}^4$, permits us to see something else that inhabitants of $\mathbb{R}^4$ have known for a long time: to every point of the real hyperbola there is attached a circle. And again, what we've seen up to now—**V** lying on the x_1- or y_1-axis—is just part of a larger picture. As before, the plane containing **V**, **W** and the circle is Minkowski.

Teacher. The analogy goes further: the plane containing the attached circle contains $\mathbf{V}$ and $\mathbf{W}$, meaning that the plane turns in $\mathbb{R}^4$, and it does so in such a way that for any two distinct values of t_1, the associated planes intersect only in the origin.

Philosopher. You'd prove that the same way you earlier proved the corresponding statement?

Teacher. Yes, the argument is entirely similar.

Philosopher. The burning question now is, how do all the circles fit together to make a sphere?

Teacher. A real hyperbola branch plays the same role as the real circle did earlier: just as then a branch passed through each point of the circle, a circle now passes through each point of the hyperbola branch. A topologically equivalent model of this is shown in Figure 12.5:

FIGURE 12.5. In the same way that Figure 12.4 is a companion to Figure 12.1, this one is a companion to Figure 12.2—covering the sphere with circles of latitude versus covering it with lines of longitude as in Figure 12.2. The longitude represents one real hyperbola branch serving as base space, and the latitudes are the various attached circles. Each circle lives in a Minkowski plane, just as the hyperbolas of Figure 12.2 do.

Philosopher. All these circles intersect the lines of longitude orthogonally, and any one of these circles is the image of a *horizontal* line in the top half of Figure 12.3.

Teacher. Yes, and notice also that since cosine and sine have period 2π, as we move along any horizontal line in complex t-space, we go around and around the circle of latitude on the sphere.

Student. I know that just plain cosine and sine are periodic, but is that true even if the horizontal line isn't the real axis in the complex t-plane, but some *other* horizontal line?

Teacher. Yes, and that's easy to show, using the addition formulas for cosine and sine. For example,

$$\cos\big((t_1+2\pi)+it_2\big) = \cos(t_1+2\pi)\cdot\cos it_2 - \sin(t_1+2\pi)\cdot\sin it_2$$
$$= \cos(t_1)\cdot\cos it_2 - \sin(t_1)\cdot\sin it_2 = \cos(t_1+it_2).$$

Similarly for sine.

Parallel Slices

Student. I once read how Flatlanders might "visualize" a sphere in 3-D: let time t be the third dimension, so the sphere's equation is, for example, $x^2+y^2+t^2=1$. If t goes from -1 to $+1$, this equation defines a "movie" of a circle growing from a point when $t=-1$ to a unit circle when $t=0$. That circle shrinks back to a point as t goes from 0 to $+1$. Couldn't we do the same to get an idea of a complex circle in 4-space?

Teacher. Yes. In your example, the sphere is called a "phase diagram" for the expanding-contracting circle. Analogously, if we choose, say, y_2 to play the role of time t, then the complex circle in $\mathbb{R}^4$ becomes a phase diagram of a changing space curve in (x_1, x_2, y_1)-space. Substituting $x = x_1 + ix_2$, $y = y_1 + it$ into $x^2+y^2=1$ and separating into real and imaginary parts gives

$$x_1^2 - x_2^2 + y_1^2 - t^2 = 1, x_1x_2 + y_1t = 0. \tag{12.13}$$

The curve at a particular time t_o, is then defined by (12.13) with $t=t_o$. When $t=0$, one gets the union of a circle and a hyperbola, a picture analogous to that in Figure 5.4 [p. 73].

For $t=t_o$, the first equation of (12.13) becomes

$$x_1^2 - x_2^2 + y_1^2 = 1 + t_o^2, \tag{12.14}$$

which defines a one-sheeted hyperboloid with the x_2-axis as axis of symmetry. For $t_o \neq 0$, the second equation is

$$y_1 = \frac{-1}{t_o}x_1x_2, \tag{12.15}$$

defining a hyperbolic paraboloid. The space curve itself satisfies both these equations, making it the intersection of the hyperboloid and paraboloid.

Philosopher. But ellipse-plus-hyperbola in Figure 5.4 is the intersection of a hyperboloid with two *planes*, not with a paraboloid.

Teacher. Those two planes are what the paraboloid becomes when $t=0$, for then the second equation in (12.13) is $x_1x_2=0$, so $x_1=0$ or $x_2=0$. $\{t=0, x_1=0\}$ gives, in Figure 5.4, the (x_1, y_1)-plane containing the ellipse; $\{t=0, x_2=0\}$ gives the (ix_2, y_1)-plane containing the hyperbola. When t_o is very small, the paraboloid looks very much like two planes, at least around the origin. In fact, if t_o is small enough, for miles around, your eyes couldn't even see any difference.

Student. Incredible! How come?

Teacher. Here's a picture of the paraboloid in (12.15) with $t_o=-1$. If we write (12.15) as

$$\frac{y_1}{\left(\frac{-1}{t_o}\right)} = x_1x_2,$$

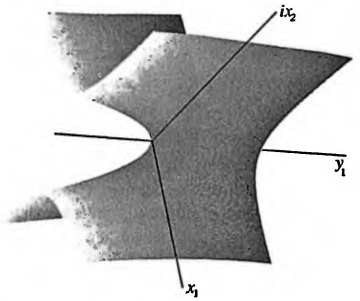


FIGURE 12.6. The hyperbolic paraboloid (saddle-surface) $y_1 = x_1x_2$. This is the equation $x_1x_2 + y_1t = 0$ in (12.13) when $t = -1$.

we see that when t_o is small, the denominator of y_1—the stretch factor in the y_1-direction—is large. As t_o decreases, the saddle surface becomes more and more like two walls at right angles, and in the limit everything reduces to two orthogonal planes.

Philosopher. So what does the space curve look like when $t_o \neq 0$?

Teacher. Here's a picture for fairly small t_o, showing the hyperboloid and stretched paraboloid intersecting in the space curve:

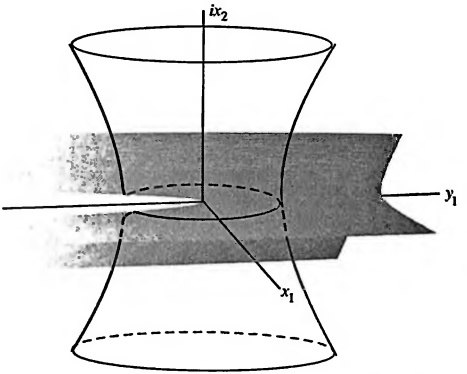


FIGURE 12.7. When $t_o < 0$ gets close to 0, the saddle surface $t_oy_1 = -x_1x_2$ is stretched out, and its intersection with the hyperboloid of one sheet $x_1^2 - x_2^2 + y_1^2 = 1 + t_o^2$ gives two space curves, as above. For t_o sufficiently small, their union looks like [circle] ∪ [hyperbola]

Notice that when an ant walks along the rightmost edge of the stretched paraboloid, he travels in roughly a straight line for a time, then makes a fairly sharp 90° turn, and finally continues in a straight line in this new direction. This approximates a path on two walls meeting orthogonally. This figure suggests looking at our circle analogue of Figure 5.4 in a new way, as the union of the solid and dashed curves:

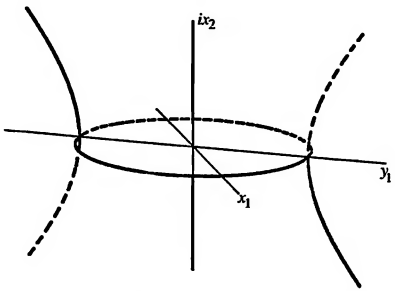


FIGURE 12.8. The picture of [ellipse] ∪ [hyperbola], which appears so often in our discussions, can be regarded as the union of the solid and dashed space curves. We can think of the solid curve's "ends" as meeting at the north pole of a sphere, and those of the dashed curve as meeting at the south pole.

Philosopher. I never once thought of our familiar union of an ellipse and hyperbola that way!

Teacher. Notice that the ends of one space curve head toward a common point at infinity; so do the end points of the other space curve, but to a different point at infinity. A little thought shows that with appropriate topological massaging, the curves for various values of t_o lie on a sphere in this way:

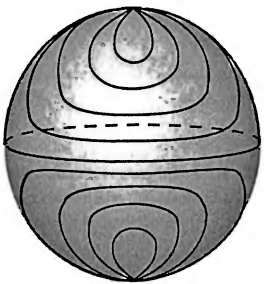


FIGURE 12.9. Space curves like the ones in Figure 12.7 form a family as t_o varies. They can be massaged to lie on a sphere in the manner shown here.

What Happens When the Circle's Radius Is 0?

Teacher. One can mimic everything we've just done when the circle's radius is 0. Then (12.13) [p. 227] becomes

$$x_1^2 - x_2^2 + y_1^2 - t^2 = 0, \quad x_1 x_2 + y_1 t = 0. \tag{12.16}$$

When $t_o = 0$, the hyperboloid becomes a cone, and the paraboloid degenerates to the two planes $x_1 = 0$, $x_2 = 0$. The space curve in (x_1, ix_2, y_1)-space is then the union of two lines through the origin:

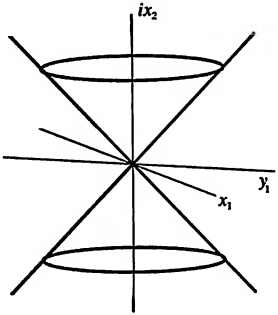


FIGURE 12.10. When both the radius and t_o are zero, the circular waist of the hyperboloid of one sheet shrinks to a point, and the hyperboloid becomes a cone. The hyperbolic paraboloid of Figure 12.7 degenerates to two planes. The cone and the plane $x_1 = 0$ intersect in two lines in the (x_2, y_1)-plane, while the cone and the plane $x_2 = 0$ intersect in just the origin—a real circle of radius 0.

Remarkably, when $t_o \neq 0$, one gets a copy of these two lines in each of the two planes $x = \pm t_o$.

Philosopher. You still get *lines*? Why?

Teacher. If $t_o \neq 0$, the second equation of (12.16) yields

$$y_1^2 = \frac{x_1^2 x_2^2}{t_o^2}$$

which, when substituted into the first, paraboloid-defining equation of (12.16), leads to $(x_1^2 - t_o^2)(x_2^2 + t_o^2) = 0$. Because all these quantities are real, only the first factor can be zero, and when it is, then $x_1^2 = t_o^2$. This, substituted into the first equation of (12.16) implies $x_2 = \pm y_1$, which gives two lines in each of the two planes $x_1 = \pm t_o$.

Philosopher. But in $\mathbb{P}^2(\mathbb{R})$, each of these lines meets the line at infinity in one point, so in every case their closure is a topological circle...

Teacher. That's right! Here is how they fit together:

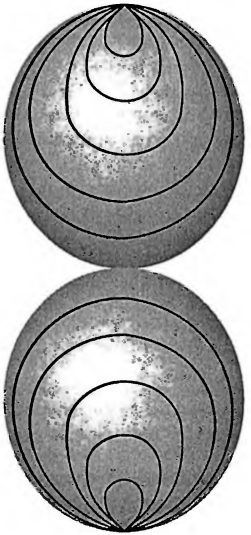


FIGURE 12.11. In Figure 12.9, squeezing the equator to a point produces the union of two spheres touching at one point. This is the topological counterpart of algebraically diminishing r to zero in $x^2 + y^2 = r^2$. In each case, when r is zero, something that was irreducible becomes reducible: one sphere becomes two; and the polynomial $x^2 + y^2 - r^2$ becomes reducible over $\mathbb{C}$, $x^2 + y^2$ factoring into $(x + iy)(x - iy)$.

Student. But that's not one sphere, but *two*! Haven't you made a mistake? How could one sphere suddenly become two?

Teacher. Let's look at the equation for a circle of radius 0, $x^2 + y^2 = 0$. This factors:

$$x^2 + y^2 = (x - iy)(x + iy) = 0.$$

As we saw earlier, if you have two loci given by $p = 0$ and $q = 0$, then $pq = 0$ gives their union. In our case, each of $x - iy = 0$ and $x + iy = 0$ defines a line.

Student. Aren't they null lines?

Teacher. Yes, but still perfectly good lines! Their slopes are $\pm i$, so $y = \pm ix$, and any point $(x, \pm ix)$ on either line has separation-squared $x^2 - x^2 = 0$.

It happens that $x^2 + y^2 - r^2$ factors only when $r = 0$.

Philosopher. You mean if $r \neq 0$, then it doesn't factor—it's "prime"? How do you know that?

Teacher. Well, suppose it *did* factor, say

$$x^2 + y^2 - r^2 = (ax + by + c)(Ax + By + C).$$

Expanding the right side and equating coefficients shows that $A = \frac{1}{a}$, $B = \frac{1}{b}$, and also that if $r \neq 0$, then c and C are nonzero. From this we see that the coefficient of xy is

$$\frac{a^2 + b^2}{ab},$$

and since that's zero, $a^2 + b^2 = 0$. There is no x- or y-term in $x^2 + y^2 - r^2$, so the coefficients of x and y there are obviously zero, and this readily leads to

$$\frac{c}{a^2} = \frac{c}{b^2}.$$

Since $c \neq 0$, this means $a^2 = b^2$ which, with $a^2 + b^2 = 0$, implies $a = 0$, $b = 0$, a contradiction.

Student. Why should we worry about all this?

Teacher. It leads to an observation I can't resist pointing out. First, there's a certain sense in which a 2-dimensional topological manifold is decomposable: intuitively, if two topological surfaces touch at only finitely many points, their union is decomposable—they aren't glued together too well, and come apart more easily than, say, a sphere. To break *that* apart, you'd have to cut it along a whole curve.

Now the two spheres in Figure 12.11 touch at just one point, so their union is in that sense decomposable, and this situation corresponds to the associated polynomial $x^2 + y^2$ being decomposable—that is, in more standard terminology, *reducible*. On the other hand the sphere isn't decomposable, and when $r \neq 0$, correspondingly $x^2 + y^2 - r^2$ isn't either. Our previous arguments easily extend to show that a complex circle of *any* nonzero radius is topologically a sphere, so in this case algebraic decomposability has a topological "picture."

Philosopher. That's a nice observation.

There's a very simple formula for the genus of any complex algebraic curve $p(x, y) = 0$, provided that as a surface in $\mathbb{P}^2(\mathbb{C})$, it has no self-intersections or cusps. If p is irreducible of degree n, then the genus is

$$g = \frac{(n-1)(n-2)}{2}.$$

For example, $p = x^n + y^n - 1$ satisfies these criteria.

Substituting in $n = 100$ gives $g = 4851$. Choosing $n = 1000$ gives $g = 498{,}501$. (See the box on p. 222.)

Conics and Sphere Coverings

Philosopher. We've seen that a complex projective circle is a topological sphere. But what about an ellipse or hyperbola?

Teacher. Even in $\mathbb{C}^2$, an ellipse is just a stretched circle, so in $\mathbb{P}^2(\mathbb{C})$ it's a topological sphere. And a hyperbola is no different from an ellipse: in $\mathbb{P}^2(\mathbb{C})$ we assume in both cases that we've included the two points at infinity, so topologically the closed complex hyperbola is a sphere, too.

Philosopher. I have another question that has been bothering me. We've convinced ourselves in several ways now that a complex circle—and therefore any nondegenerate complex projective conic—is a topological sphere. But I've seen books showing how its equation defines a double covering of the Riemann sphere with exactly two branch points, from which one easily concludes that we have a topological sphere.

Teacher. Yes, other than at those x's where we have repeated roots, the equation

$$Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0$$

has two distinct y-values, and therefore defines in a natural way a 2-covering of the Riemann sphere. Those repeated roots happen to occur for two values of x, and can define the endpoints of two "cuts" along which we switch freshly-cut edges of two concentric spheres to make a topological sphere.

Philosopher. That approach is spelled out in many books, and not only gives the topology of a nondegenerate complex conic, it even extends to higher degrees, where we can get a genus > 0. So what's the advantage of all our slicing?

Teacher. Our curves in 3-slices give us a view of part of the complex circle that is *undistorted* in $\mathbb{R}^4$. Sphere coverings don't give us this kind of information. A circle curves, and the complex circle, viewed in $\mathbb{R}^4$, has curvature, also. The real circle suggests that the curvature of a complex circle is probably constant, too, but our limited, 2-D view so easily leads to wrong assumptions! Remember, at every point of the real circle, there attaches a hyperbola branch, and very far out, when these branches get very close to their asymptotes, the curvature decreases.

Student. This seems exciting—the prospect of discovering that the curvature of a complex circle is not constant! How do we do that?

Chapter Summary

When we extend $\mathbb{R}^2$ to $\mathbb{P}^2(\mathbb{C})$, conics in $\mathbb{R}^2$ grow, too, and it's natural to ask what their extensions look like. In this chapter, we find that any ellipse, hyperbola or parabola extends to a topological sphere. Although this can be readily established using a standard sphere-covering argument, that quick route comes at the price of introducing metric distortion. This chapter instead presents two methods of seeing how a complex circle naturally lies in $\mathbb{C}^2$. Each of these two dual methods shows that a circle or any other nondegenerate complex conic must have nontrivial curvature that is much more informative than the real-curve counterpart. This leads to the next chapter on Gauss curvature.

Exercises

1. Carry out the derivation suggested in the box on p. 224 about addition formulas for hyperbolic functions.
2. Show that as each of $\mathbf{v}$ and $\mathbf{w}$ in Figure 12.1 [p. 219] rotates around a circle, the endpoint of $\mathbf{v}+\mathbf{w}$ always lies on a certain sphere S^3 in $\mathbb{R}^4$. What is its radius?
3. Do the hyperbola branches in Figure 12.1 actually attach to the circle "at right angles"? One can experiment: pick, say, the point

$$P=\left(\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}}\right)$$

on the circle shown there, and find the equation of the line tangent to the circle at P. Now translate that line a tiny bit this way and that, letting Maple find the points of intersection with the complex circle. This gives various vectors from P to points on the complex circle. It's clear what the vectors to points of the *real* circle look like. But what about the vectors to points of the attached hyperbola? Are they, in the limit, perpendicular to the circle vectors? Perform the map

$$(x_1, ix_2, y_1, iy_2) \to (x_1, x_2, y_1, y_2)$$

from $\mathbb{C}^2$ to $\mathbb{R}^4$, and find the angle between the real vectors using the dot-product. The angle should approach 90° as the amount of translation each way goes to 0. Try this idea on some other points of the real circle less "arbitrary" than the above P.

4. Give a short argument that the completion in $\mathbb{P}^2(\mathbb{C})$ of the union of any two parallel complex lines in $\mathbb{C}^2$ is topologically the same as a complex circle of zero radius.
5. Explain how to draw a topological picture of the completion in $\mathbb{P}^2(\mathbb{C})$ of n mutually parallel complex lines in $\mathbb{C}^2$.
6. Find a polynomial equation $p(x, y) = 0$ topologically defining, in $\mathbb{P}^2(\mathbb{C})$:
   (a) Two spheres touching at two different points;
   (b) Three spheres touching at three different points.
   (c) Is there a p producing two spheres touching at three points?
   (d) Is there a p producing three spheres touching at two points?

13

Curvature

in which our trio studies
the curvature of surfaces in $\mathbb{R}^3$
in preparation for still mightier things

Curvature of Curves

Philosopher. I've been thinking... our everyday intuition about conics is based on the appearance of only their real portions. In that customary and not-very-representative slice of $\mathbb{C}^2$, we see just one pair of foci, so everyone thinks ellipses or hyperbolas have just one pair. We know now that they actually have two pairs, one on each principal axis. Similarly, looking only in the real plane, we never see the imaginary area of a hyperbola. We also see just a little part of the sum-of-lengths story. And so on. The "larger truths" we've been discovering seem more satisfying, consistent and beautiful—even if sometimes things are more subtle.

Teacher. I agree.

Philosopher. This 3-D slice of a complex circle strongly suggests that perhaps yet another "icon" must fall.

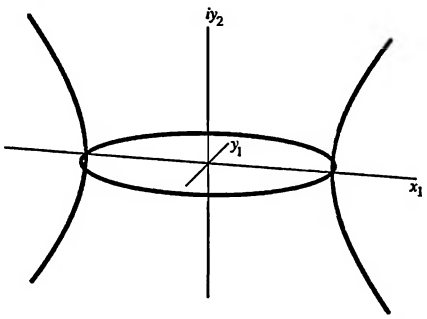


FIGURE 13.1. A real circle has constant curvature. But this 3-D slice of the complex circle in $\mathbb{R}^4$ raises suspicions—the curvature of the attached hyperbola is not at all constant!

Student. Always looking at the *real* circle has made us fail for another hoax!

Philosopher. Admittedly, the real circle does have the same curvature at each point.

Student. What do you mean by "curvature at a point"?

Philosopher. Roughly, the amount the curve turns per small amount of travel near the point. For example, when you drive a car around a bend, the tighter the bend, the greater the amount of turning the car does per 5 feet, say. If the car travels in a circle, then from a circle's symmetric nature, one sees that the car's light beams turn the same amount for a given distance traveled. It is in this sense that the real circle has constant curvature. But drive along a parabola, and you find that the car does most of its turning near the vertex. Further out, the path gets straighter—the curvature there is less.

Student. That seems reasonable.

Philosopher. The same thing happens on a hyperbola branch: it curves the most near its vertex, but straightens out as it approaches an asymptote. Because of this, when I look at Figure 13.1, I see an icon in peril. It seems that the full complex circle in $\mathbb{R}^4$ can't actually have constant curvature!

Teacher. You raise an intriguing issue. However, a complex circle is really a *surface*, so the actual question is, how much does that *surface* curve at various points?

Student. Not only that, but the surface sits in 4-space! I sense a gathering storm that could turn into a hurricane!

Teacher. Possibly—but let's just see where the truth leads us! I suggest we begin with curves and build up to surfaces. Your idea of curvature being the amount of turning per unit length is very typical of how we measure all sorts of "per" quantities. Take mph—miles per hour, for example. A car traveling down a road will have slightly varying speeds from instant to instant, but the distance covered in a particular microsecond, divided by that microsecond, gives a good approximation to the speed during that short time interval. We can define the theoretical speed as the limit taken over ever-shorter time intervals. One could in the same spirit measure the amount of fuel consumed as the front wheels make one revolution, thus getting a real-time approximation to mpg—miles per gallon.

Student. What if cars had an mpg dial next to the usual mph dial! I bet seeing that needle constantly changing would get a person's attention. I mean, that's *money*!

Teacher. It would certainly provide feedback, wouldn't it? Drivers might even unconsciously train themselves to reduce fuel waste! Anyway, in place of mph or mpg, one measure of curvature might be, say, "rpf": radians per foot—the number of radians the car turns as it travels a foot.

Student. That means the curvature is almost always really small, doesn't it? That is, going five feet down a highway, a car would turn just a fraction of a degree, and therefore an even smaller fraction of a radian!

Teacher. That is right. But generally, a "per" measurement depends upon units and one can choose units convenient to any particular application. Sixty mph is 1 mile per minute

and a "mere" $\frac{1}{60}$ miles per second. Miles per second might be OK for measuring the speed of light, but it would be pretty unusual to use that to measure how fast a car is going. In measuring the curvature of the earth, degrees per 100 miles might be reasonable, while radians per micrometer might be appropriate for measuring the curvature of a dot of ink from an inkjet printer.

Philosopher. What would be a mathematical definition of curvature of a curve at a point?

Teacher. Well, if s denotes arc length measured from some point on a smooth curve C in $\mathbb{R}^2$ (see Appendix 2), and if $\theta = \theta(s)$ measures the corresponding signed angle the tangent line makes with the horizontal, then one can define the curvature κ to be $\frac{d\theta}{ds}$. That is, if a point $P \in C$ corresponds to s_o, then the curvature of C at P is $\frac{d\theta}{ds}$ evaluated at s_o.

> $\frac{d\theta}{ds}$ is more precisely the *signed* curvature, and that's what we use. The *unsigned* curvature $\left|\frac{d\theta}{ds}\right|$ is useful if you don't care whether the curve veers to the right or left.

Philosopher. This is all very simple and elegant. Actually, I see no compelling reason why it must be the tangent line. Couldn't one just as well use, say, a *normal* to the plane curve? After all, being perpendicular to the tangent, the normal turns just as much as the tangent does!

Teacher. You're absolutely right—one could just as well use a normal direction.

Student. How would one use this definition to calculate a specific curvature? Can you give me an example?

Teacher. An easy and natural example is the circle $x^2 + y^2 = r^2$. The angle θ from the origin specifies, say, the direction of the outward direction of the normal, and if θ is measured in radians, then $s = r\theta$. Of course that means $ds = r\,d\theta$—that is,

$$\frac{d\theta}{ds} = \frac{1}{r}.$$

For a circle of radius r, we therefore have $\kappa = \frac{1}{r}$.

> The numerical value of curvature depends in an essential way on the units being used. The curvature of the earth's equator is small if you measure arc length in feet or miles, but if you use light-years, the number is *tremendous*. At such a cosmic scale, we're just a tiny round point of light—like a dot of ink from an inkjet printer on a vast sheet of paper.

Student. Can you actually use your definition to get the curvature of an ellipse, for example?

Teacher. Yes. Let's begin with an "easy" point—a vertex. To do this for the ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1,$$

I'm going to pick a point P on the ellipse close to $(a, 0)$. If P has y-coordinate $b\epsilon$ (where ϵ is small and positive), then

$$\frac{x^2}{a^2} + \frac{(b\epsilon)^2}{b^2} = 1$$

leads to $x = a\sqrt{1-\epsilon^2}$. In Figure 13.2, the length along the curve from $(a, 0)$ to P is approximately the length of the heavily-drawn vertical line segment. That is, $\Delta s \approx b\epsilon$.

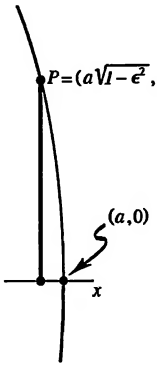


FIGURE 13.2. For ϵ small, the distance traveled along the ellipse from $(a, 0)$ to P is approximately $b\epsilon$. We find that the corresponding angle $\Delta\theta$ through which the normal to the ellipse turns is approximately $\frac{a\epsilon}{b}$. From these facts we conclude that the ellipse has curvature $\frac{a}{b^2}$ at $(a, 0)$.

Now in going from $(a, 0)$ to P, how much turning $\Delta\theta$ has, say, the outward ellipse normal done? Such a normal vector is given by the gradient of

$$\frac{x^2}{a^2} + \frac{y^2}{b^2}.$$

This is

$$\left(\frac{2x}{a^2}, \frac{2y}{b^2}\right),$$

which at P simplifies to

$$\left(\frac{2\sqrt{1-\epsilon^2}}{a}, \frac{2\epsilon}{b}\right).$$

Therefore

$$\Delta\theta \approx \tan\Delta\theta = \frac{\frac{2\epsilon}{b}}{\frac{2\sqrt{1-\epsilon^2}}{a}} = \frac{\frac{\epsilon}{b}}{\frac{\sqrt{1-\epsilon^2}}{a}}$$

Now we saw just a moment ago that $\Delta s \approx b\epsilon$, so dividing through by Δs gives

$$\frac{\Delta\theta}{\Delta s} \approx \frac{\frac{1}{b^2}}{\frac{\sqrt{1-\epsilon^2}}{a}}.$$

As $\epsilon \to 0$, the limit of this ratio is $\frac{a}{b^2}$, so the curvature of the ellipse at $(a, 0)$ is

$$\kappa = \frac{a}{b^2}. \tag{13.1}$$

Philosopher. I guess that by symmetry, the curvature at the vertex $(0, b)$ is $\kappa = \frac{b}{a^2}$.

Teacher. That's right.

Philosopher. It would be nice to check that these answers make sense. For a 2-by-1 ellipse, the formula makes the curvature κ *eight* times larger at $(a, 0)$ than at $(0, b)$. That sounds somewhat surprising to me. Here's the picture:

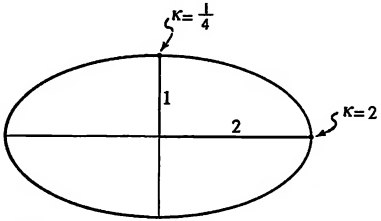


FIGURE 13.3. The ratio

$$\frac{a}{b^2} \div \frac{b}{a^2} = \left(\frac{a}{b}\right)^3$$

between extreme curvatures on a real ellipse can get large surprisingly quickly when $a > b$—in this case, 8 for $a = 2, b = 1$.

I'm now wondering whether this picture corroborates, or casts doubts, on our formula.

Teacher. It turns out that the curvature of a curve at a point is equal to the curvature of the circle of best fit at that point. In Figure 13.4, next, I have drawn the closest-fitting circle to the ellipse at (2,0) and also the one at (0, 1).

Interactive Applet

For an interactive version of Figure 13.4, see Chapter Applet 18 on the CD-ROM.

For a real ellipse ten times longer than it is wide, the ratio of the maximum to the minimum curvature is 1000.

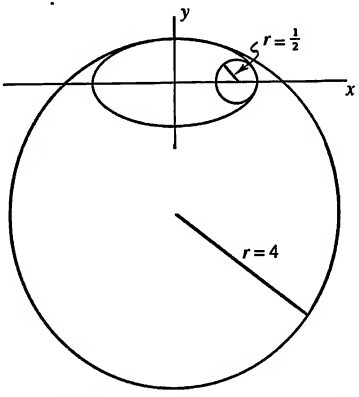


FIGURE 13.4. At the top of the ellipse, the relatively gentle turning makes for quite a large best-fit circle. At the rightmost point of the ellipse, the tighter turning means a smaller circle. In this case, the radii are in an 8:1 ratio.

The circle at (2,0) has curvature $\kappa = 2$, meaning it has radius $\frac{1}{2}$. The one at (0, 1) has curvature $\frac{1}{4}$, or radius 4. This agrees nicely with what our formula predicts.

Philosopher. You used the term "circle of best fit at a point." Visually, they fit well, but is there a precise definition of this?

Teacher. Yes. Let P, P' be close-together but distinct points on a smooth curve in $\mathbb{R}^2$. [See Appendix 2.] Let L, L' be the respective normal lines through these points, and let $C(P, P')$ be the circle through P and centered at $L \cap L'$. If c is the limiting location of $L \cap L'$ as $P \to P'$, then the circle of best fit at P is the circle through P with center c.

Student. Formula (13.1) seems nice, but can your definition lead to a more general formula valid at *any* point of an ellipse?

Teacher. Actually, one can do even better than that—there's a formula valid at an arbitrary point of a smooth curve in $\mathbb{R}^2$. And it's not hard to see that an ellipse is smooth.

> In discovering the differential calculus, one of the first things Newton (1642–1727) worked out, besides the slope of a tangent line, was the curvature of a parametrically-defined curve.

Philosopher. Well, I know an ellipse has parametrization $x = a\cos t$, $y = b\sin t$. Can we use this parametrization to get a formula for the curvature at any point of an ellipse?

Teacher. Yes. Let $x = x(t)$, $y = y(t)$ be a smooth parametrization of a smooth curve C. We can arrive at such a curvature formula simply by converting the geometrically-

inspired definition of curvature, $\frac{d\theta}{ds}$, into rectangular coordinates—that is, by expressing the curvature in terms of $x(t)$ and $y(t)$.

We can write

$$\frac{d\theta}{ds} \quad \text{as} \quad \frac{\frac{d\theta}{dt}}{\frac{ds}{dt}},$$

or in dot-notation, as $\frac{\dot{\theta}}{\dot{s}}$. So we will set about expressing $\dot{\theta}$ and $\dot{s}$ in terms of x and y. Now the angle θ between a tangent ray to C and the positive x-axis satisfies

$$\tan\theta = \frac{dy}{dx} = \frac{\dot{y}}{\dot{x}}, \quad \text{so} \quad \theta = \text{Arctan}\left(\frac{\dot{y}}{\dot{x}}\right).$$

Furthermore, the differentiation formula

$$\frac{d}{dt}(\text{Arctan}(u)) = \frac{\dot{u}}{1+u^2}$$

means that in our case,

$$\dot{\theta} = \frac{d}{dt}\left(\text{Arctan}\left(\frac{\dot{y}}{\dot{x}}\right)\right) = \frac{1}{1+\left(\frac{\dot{y}}{\dot{x}}\right)^2} \cdot \frac{d}{dt}\left(\frac{\dot{y}}{\dot{x}}\right),$$

which works out to

$$\dot{\theta} = \frac{\dot{x}\ddot{y} - \ddot{x}\dot{y}}{(\dot{x})^2 + (\dot{y})^2}.$$

Next, to get $\dot{s}$ in terms of x and y, write $\mathbf{V}(t) = (x(t), y(t))$, and look at this picture:

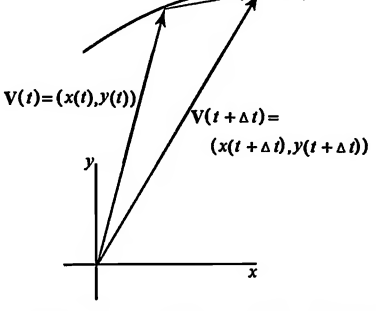


FIGURE 13.5. The difference vector $\Delta\mathbf{V} = (x(t+\Delta t)-x(t), y(t+\Delta t)-y(t))$; the limit of $\frac{\Delta\mathbf{V}}{\Delta t}$ as $\Delta t \to 0$ is $(\dot{x}(t), \dot{y}(t))$.

From it, we see that the vector $\frac{\Delta \mathbf{V}}{\Delta t}$ is

$$\left(\frac{x(t+\Delta t)-x(t)}{\Delta t}, \frac{y(t+\Delta t)-y(t)}{\Delta t}\right).$$

Taking the limit of this as $\Delta t \to 0$ shows that the vector $\frac{d\mathbf{V}}{dt}$ is $(\dot{x}(t), \dot{y}(t))$, and that in turn means that the length of this vector is

$$\dot{s} = \sqrt{\frac{d\mathbf{V}}{dt} \cdot \frac{d\mathbf{V}}{dt}} = \sqrt{\dot{x}^2(t) + \dot{y}^2(t)}.$$

Therefore

$$\frac{\dot{\theta}}{\dot{s}} = \frac{\dot{x}\ddot{y} - \ddot{x}\dot{y}}{((\dot{x})^2 + (\dot{y})^2)^{3/2}}.$$

Since $\frac{\dot{\theta}}{\dot{s}} = \frac{d\theta}{ds} = \kappa$, we have our formula:

$$\kappa = \frac{\dot{x}\ddot{y} - \ddot{x}\dot{y}}{((\dot{x})^2 + (\dot{y})^2)^{3/2}}. \tag{13.2}$$

Student. That looks sort of familiar. In my calculus book, there's a formula for the curvature of the graph of the function $y = y(x)$.

Teacher. Yes. That curvature formula is

$$\kappa = \frac{\ddot{y}}{(1 + (\dot{y})^2)^{3/2}}, \tag{13.3}$$

and it's a special case of (13.2). That is because an ordinary function $y = y(x)$ can be written in parametric form as $x = t$; $y = y(t)$, so $\dot{x} = 1$, and $\ddot{x} = 0$. Substituting these into (13.2) yields (13.3).

It's well known that the second derivative at a point measures how convex or concave the graph of a function $y = f(x)$ is at that point. Specifically,

$$\frac{d^2 f}{dx^2}$$

measures the rate of change of slope with respect to x. In contrast, curvature measures the rate of change of *angle* with respect to *arc length*. However, at a critical point (where $f' = 0$), the two are the same. This is because on the one hand, $\Delta x \approx \Delta s$ there; and on the other hand, the slope $\tan(\theta)$ is small near a critical point, meaning that $\tan(\theta) \approx \theta$ and therefore $\Delta(\tan(\theta)) \approx \Delta\theta$ there. At such a point, (13.3) simplifies to $\kappa = f''$, and we immediately know such things as: the parabola $y = x^2$ has curvature $\kappa = 2$ at (0, 0), the graph of $y = \cos(x)$ or $y = \sin(x)$ has $\kappa = \pm 1$ at each of its maxima and minima, and for any $n > 2$, the graph of $y = x^n$ has zero curvature at (0,0).

Philosopher. When I compare the general formula (13.2) with the special one in (13.3), it seems that the general one is both more symmetric and more beautiful.

Student. That new formula also seems powerful. To get the curvature of an ellipse at any point, you'd just use the parametrizations $x = a\cos t$, $y = b\sin t$, wouldn't you?

Teacher. Exactly. Those parametric equations lead to

$$\dot{x}(t) = -a\sin t;\ \dot{y} = b\cos t;$$

$$\ddot{x}(t) = -a\cos t;\ \ddot{y} = -b\sin t.$$

Substituting these into (13.1) and simplifying a bit gives

$$\kappa_{\text{ellipse}} = \frac{ab}{(a^2\sin^2 t + b^2\cos^2 t)^{3/2}}. \tag{13.4}$$

Philosopher. This simplifies so nicely to the curvatures at the vertices! When $t = 0$, the parametrization puts us at $(a, 0)$, and (13.4) reduces to $\frac{a}{b^2}$. When $t = \frac{\pi}{2}$, we're at $(0, b)$, and (13.4) produces $\frac{b}{a^2}$.

Teacher. Yes. In fact, one can use (13.4) to plot a graph, showing how the curvature of our 2-by-1 ellipse smoothly varies between the extremes attained at the vertices:

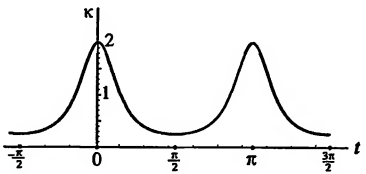


FIGURE 13.6. The 2-by-1 ellipse parametrized by $x = 2\cos t$, $y = \sin t$, has curvature at (x, y) given by

$$\kappa = \frac{2}{(1 + 3\sin^2 t)^{1.5}},$$

which is what (13.4) becomes when $a = 2$, $b = 1$. The extreme curvatures are 2 and $\frac{1}{4}$.

Philosopher. There is another nice corroborating check for our formulas: they're all dimensionally correct.

Student. What do you mean?

Philosopher. In physics, one traditionally uses L, T and M to represent the classical dimensions of length, time and mass, and other classical quantities can be expressed in terms of these three. In our case, curvature is change in angle per unit length, and angles are dimensionless. Therefore $\frac{d\theta}{ds} = \kappa$ must have dimension L^{-1}. This is in fact true for a circle, since its curvature is $\frac{1}{r}$, which has dimension L^{-1}. What I like about (13.4) is that its dimension is also L^{-1}.

Student. Angle measurement is dimensionless? I had a teacher once that marked off points if I left off the degrees symbol! Why do you say θ has no dimension?

Teacher. Because in measuring angles, there is a natural, "universal" reference—a circle of any size. Thought of this way, an angle is simply the *percentage of a circle.* From there, one can introduce various scaling factors to get a variety of convenient units, just as one can change from feet to inches or miles. So 100% of a whole turn is 360 degrees, the number 360 having evolved centuries ago because of its computational convenience: numerically 360 can be evenly divided in a lot of different ways—into 2, 3, 4, 5, 6, 9, 10, 12, 15, 18, 20, 24, 30...parts. Of course 100% of a whole turn is also 2π radians, where the angle of an arc, measured in radians, is the arc's length divided by the radius of the circle. Notice that dimensionally, this is just $\frac{L}{L} = 1$. Radians are natural, since for a unit circle, arc length is then numerically the same as the central angle measured in radians. Also, radians are the unit of angle measurement making differentiation formulas for trigonometric functions the simplest. So your teacher didn't actually mark down for leaving off a dimension, but rather for not specifying the size of angle-unit.

Philosopher. Probabilities form another example! All probabilities are pure, dimensionless quantities. When we say "the odds are 3 to 2," we're rescaling probabilities $\frac{3}{5}$ and $\frac{2}{5}$ by multiplying by a common denominator to make the measurements come out whole numbers.

Teacher. Sure. It's important to recognize that, just as changing the unit of length affects the number we call curvature, so does angle measurement. A highway engineer might think of a road as curving 5 degrees per 100 feet; that number would be different, and probably more convenient for his purposes, than measuring in radians or whole turns per 100 feet.

Surface Curvature

Philosopher. When it comes to measuring curvature of a surface, how can we build upon what we've done for curves? On a curve, we measured the limiting amount of turning per unit length traveled along the curve. But from any point on a surface, one can travel away on the surface in many different directions.

Teacher. We can try higher-dimensional analogues. That is, instead of measuring per unit length, try measuring per unit area. And angles? There's a higher-dimension analogue for them, too—solid angles!

> In analogy to an ordinary angle being a percentage of a circle, a solid angle can be regarded as a percentage of a sphere. As with ordinary angles, solid angles are dimensionless. In analogy to the radian, there's a natural solid-angle measure, the *steradian.* It's well-named, for it is essentially a stereoscopic, or 3-D radian: just as one radian subtends a unit arc length on the unit circle, the steradian subtends a unit area on the unit sphere. So just as 100% of a circle is $2\pi 1 = 2\pi$ radians, 100% of a sphere is $4\pi 1^2 = 4\pi$ steradians.

Philosopher. How do you propose using solid angles to measure curvature of a surface?

Teacher. Your earlier observation about using normals instead of tangents for measuring angles comes in handy here. At each point P of a smooth surface in $\mathbb{R}^3$ defined, say, by $f(x, y, x) = 0$, there's a well-defined normal vector given by the gradient ∇f evaluated at P. If $\nabla f \neq 0$, we can normalize it, thus obtaining a point of the unit sphere. As one "paints" a small region of area ΔA on the surface, the corresponding unit vectors paint a corresponding set of area $\Delta\Theta$ on the unit sphere. This correspondence from the surface to the unit sphere has been named the *Gauss map*. Gauss used this map extensively in his famous 1827 paper on differential geometry, *Disquisitiones generales circa superficies curvas*.

In this remarkable paper, he takes considerable care to prove that the ratio $\frac{\Delta\Theta}{\Delta A}$ actually *does* approach a limit $\frac{d\Theta}{dA}$ as the area ΔA approaches zero—that is, no matter how the region ΔA shrinks to the point P, the limiting ratio is well-defined. This is a very nontrivial fact, because there are many different ways the little region can shrink to a point. For example, it might stay tall and skinny, or stay short and wide, or remain roughly circular, or vary through a range of shapes. *This well-defined limit is taken to be the size of the curvature of the surface at P*

Now "size" has the connotation of being nonnegative. But Gauss realized something more: *nonzero curvature can in a natural way be assigned a positive or negative sign.* It turns out that the nature of curvature at point of an ellipsoid is in a sense opposite to that at any point of a saddle surface. The essential difference is revealed as we apply the Gauss map to a small oriented loop drawn on either surface: if the loop lies on an ellipsoid, then the Gauss map *preserves* the loop's orientation—that is, if we draw a little loop on the ellipsoid by circling some point counterclockwise, then its image on the unit sphere is likewise traced out counterclockwise. On the other hand, if the loop is drawn counterclockwise on a *saddle surface*, then the image on the unit sphere is traced out *clockwise*—the Gauss map *reverses* orientation. At an orientation-preserving point, the curvature is taken to be positive, and at an orientation-reversing point, it's taken to be negative.

It is not too hard to visualize the Gauss map preserving orientation for an ellipsoid, and certainly when the ellipsoid is a sphere, that is completely clear. On a saddle surface, however, things are not as obvious. A specific example provides the flavor of things at a saddle point. Let's look at a region around the point $(0, 0, 0)$ of the surface $z = y^2 - x^2$. This is a hyperbolic paraboloid—the "horse's saddle" often seen in calculus texts. One can get the coordinates of a unit normal vector at any of its points using the gradient: write the surface equation as $F(x, y, z) = x^2 - y^2 + z = 0$; then a normal is $\nabla F = (2x, -2y, 1)$, which can easily be normalized. Now in $\mathbb{R}^3$'s (x, y)-plane, let S be the square in the (x, y)-plane having the four vertices $(\pm 1, \pm 1, 0)$, and suppose we make one counterclockwise trip around the loop on the saddle surface lying directly above S. Start at the SE corner $(1, -1, 0)$. A unit surface normal there is obtained from $\nabla F(1, -1, 0) = (2, 2, 1)$ by normalizing, is $\left(\frac{2}{3}, \frac{2}{3}, \frac{1}{3}\right)$. This is the NE corner of the image loop, a "bent square" S′ on the sphere. As we proceed north along the east side of S, only the y-coordinate of S varies, and it increases from -1 to 1. The formula $\nabla F = (2x, -2y, 1)$ tells us that during this leg of the trip, only the y-component of our normal varies, and it *decreases* (from $\frac{2}{3}$ to $-\frac{2}{3}$). So already we see the Gauss-map image starting its trip in the opposite sense. One can continue around all four sides of S, the entire trip covering in turn sides E, N, W, S, while in the image loop

S′, sides are respectively encountered in the order E, S, W, N. We thus see that the image of S′ is opposite to that of S.

We therefore assign a "+" to the curvature's magnitude in the ellipsoid case, and a "−" to its magnitude in the saddle-surface case.

Student. Since solid angles are dimensionless and area is L^2, I suppose the dimension of surface curvature is L^{-2}?

Philosopher. That's right.

> Suppose that the sides of a small rectangle drawn on a sphere of radius r subtend small central angles θ_1 and θ_2. The rectangle's area is then approximately $(r\theta_1)\cdot(r\theta_2)$. If we approximate the solid angle Θ by $\theta_1\cdot\theta_2$, then we see that this area is approximately $r^2\Theta$; $r^2\Theta$ is the area of a shape on the sphere subtending a central solid angle of exactly Θ, making $r^2\Theta$ a two-dimensional analogue of the arc length $r\theta$ on a circle of radius r.

In fact, we can easily see what the curvature of a sphere of radius r is, in steradians per unit of area. Let's work in analogy to finding the curvature (measured in radians per unit arc length) of a circle of radius r: the area on a sphere of radius r having solid angle Θ steradians is $A = r^2\Theta$, so the ratio of solid angle to area is

$$\frac{\Theta}{r^2\Theta} = \frac{1}{r^2}.$$

Student. In analogy to what we did for an ellipse, can we use this definition to find the curvature at some point of an ellipsoid?

Teacher. Yes, and we can use the same general approach, too. Suppose the ellipsoid is

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1.$$

In the ellipse case, we chose a point with $x = a\sqrt{1-\epsilon^2}$ near $(a, 0)$. Let's do the same here: choosing $x = a\sqrt{1-\epsilon^2}$ near $(a, 0, 0)$ means the plane $x = a\sqrt{1-\epsilon^2}$ intersects the ellipsoid in a small ellipse

$$\frac{\left(a\sqrt{1-\epsilon^2}\right)^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1,$$

which simplifies to

$$\frac{y^2}{(\epsilon b)^2} + \frac{z^2}{(\epsilon c)^2} = 1. \tag{13.5}$$

The area of this ellipse is $\pi\epsilon^2 bc$. A parametrization of it in $\mathbb{R}^3$ is

$$\begin{aligned} x &= a\sqrt{1-\epsilon^2}; \\ y &= \epsilon b\cos t; \\ z &= \epsilon c\sin t. \end{aligned} \tag{13.6}$$

Now let's find the area on the unit sphere painted by the unit normal. The gradient of

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}+\frac{z^2}{c^2}$$

is the vector

$$\left(\frac{2x}{a^2},\frac{2y}{b^2},\frac{2z}{c^2}\right),$$

which is normal to the ellipsoid. Substituting (13.6) into this gradient gives

$$2\left(\frac{\sqrt{1-\epsilon^2}}{a},\frac{\epsilon\cos t}{b},\frac{\epsilon\sin t}{c}\right).$$

Multiplying through by $\frac{a}{2}$ "almost" normalizes it:

$$\left(\sqrt{1-\epsilon^2},\frac{\epsilon a\cos t}{b},\frac{\epsilon a\sin t}{c}\right).$$

In fact, as $\epsilon \to 0$, this vector approaches $(1,0,0)$ on the unit sphere. For $\epsilon > 0$, the parametrized curve defined by the last two coordinates of this vector lies in the plane $x=\sqrt{1-\epsilon^2}$, forming an ellipse with area of magnitude

$$\Theta=\frac{\pi\epsilon^2a^2}{bc}.$$

Since the corresponding area of the small ellipse parametrized by (13.6) is $A=\pi\epsilon^2bc$, their ratio is

$$\frac{\Theta}{A}=\left(\frac{\pi\epsilon^2a^2}{bc}\right)\Big/(\pi\epsilon^2bc)=\frac{a^2}{b^2c^2}. \tag{13.7}$$

Philosopher. I assume that as with an ellipse, one can by symmetry write down curvatures for the other vertex points of an ellipsoid. That is, at $(0,b,0)$, the curvature is

$$\frac{b^2}{a^2c^2},$$

and at $(0,0,c)$, it's

$$\frac{c^2}{a^2b^2}.$$

Teacher. Yes.

Philosopher. I believe I've just noticed something! Look at this picture:

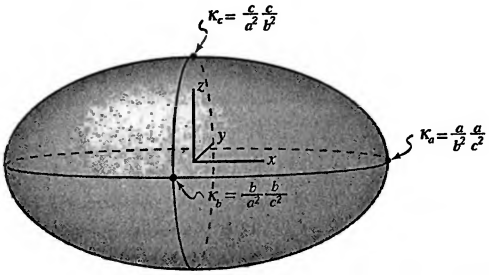


FIGURE 13.7. The Gaussian curvature at any vertex point P of the ellipsoid

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}+\frac{z^2}{c^2}=1$$

is the product of the curvatures at P of the two "coordinate ellipses" through P.

There are two coordinate ellipses through $(a, 0, 0)$, one lying in the (x, y)-plane and the other in the (x, z)-plane. *The curvature of the ellipsoid at the point* $(a, 0, 0)$ *is just the product of the curvatures of these two ellipses there*:

$$\frac{a^2}{b^2c^2}=\frac{a}{b^2}\cdot\frac{a}{c^2}.$$

Teacher. That *is* remarkable, isn't it?

Philosopher. A similar thing holds at the other vertex points, too.

Teacher. The observation that

$$\frac{a^2}{b^2c^2}=\frac{a}{b^2}\cdot\frac{a}{c^2},$$

though algebraically trivial, seems to have larger consequences because of its geometrical content. In our definition of surface curvature $\frac{d\Theta}{dA}$, both the numerator $d\Theta$ and denominator dA define "higher-dimensional analogues" of $d\theta$ and ds, and in a sense, at any vertex point each of $d\Theta$ and dA factors into the two two-dimensional parts corresponding to the coordinate ellipses through that vertex. One can write, in self-explanatory notation,

$$\frac{d\Theta}{dA}=\frac{d\theta_1}{ds_1}\cdot\frac{d\theta_2}{ds_2}. \tag{13.8}$$

Philosopher. As happens so often in science and mathematics, breaking things down into smaller constituents or building blocks can lead to progress.

Teacher. Let's concentrate for a moment on the vertex $(a, 0, 0)$. If you spin a plane containing the x-axis about the x-axis, this spinning plane intersects the ellipsoid in a family of ellipses, each one having some curvature at $(a, 0, 0)$. If $b \neq c$, then of all those ellipses, the two you mentioned are the ones having the two *extreme* curvatures—the greatest and the least.

Philosopher. Why is that?

Teacher. Let $x = u$, $y = (\cos\phi)v$ and $z = (\sin\phi)v$ parametrize a typical one of these planes. Substituting these equations into our basic ellipsoid equation

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1,$$

and putting the resulting ellipse into standard form, gives

$$\frac{u^2}{a^2} + \frac{v^2}{\left(\frac{cb}{\sqrt{(c\cos\phi)^2+(b\sin\phi)^2}}\right)^2} = 1.$$

Then (13.1) [p. 239] shows that the curvature of this ellipse at $(a, 0, 0)$ is

$$\kappa = \frac{a((c\cos\phi)^2 + (b\sin\phi)^2)}{c^2b^2}. \tag{13.9}$$

This attains its extrema where $f(\phi) = (c\cos\phi)^2 + (b\sin\phi)^2$ does. Now we can rewrite $f(\phi)$ as $b^2 + (c^2 - b^2)\cos^2\phi$. Since we're assuming $b \neq c$, this has extrema given by $\phi = 0$ or $\phi = \frac{\pi}{2}$, which correspond to the planes $z = 0$ and $y = 0$.

Student. Can you show me a picture of the curvature κ as ϕ varies?

Teacher. Sure—here it is:

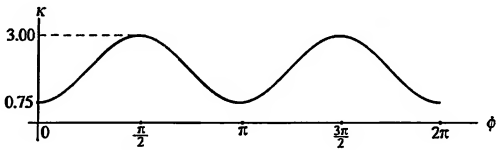


FIGURE 13.8. If one spins a plane containing the normal at a point P of a smooth surface, the rotating planes intersect the surface in curves whose curvature at P varies like a translated sine curve. Here the surface is the ellipsoid

$$\frac{x^2}{3^2} + \frac{y^2}{2^2} + \frac{z^2}{1^2} = 1,$$

and $P = (3, 0, 0)$. Since the extreme curvatures correspond to $\phi = 0$ and $\frac{\pi}{2}$, (13.9) shows that the extreme curvatures are

$$\frac{a}{b^2} = \frac{3}{4} \quad \text{and} \quad \frac{a}{c^2} = \frac{3}{1}.$$

This agrees with what we see in the graph.

Seeing "dA" and "$d\Theta$" is also revealing. For $b = 4$ and $c = 1$, Figure 13.9 shows

$$\frac{y^2}{(\epsilon b)^2} + \frac{z^2}{(\epsilon c)^2} = 1$$

(the solidly-drawn ellipse) and, superimposed for easy comparison, a few dashed elliptical images that the normals paint on the unit sphere. The smallest of these dashed ellipses corresponds to $b = 4$, $c = 1$, and $a = 1$. The successively-larger dashed image ellipses correspond to $b = 4$, $c = 1$ and $a = 2, 3, 4, 5$. The ellipsoid's curvature at $(a, 0, 0)$ increases as the square of a.

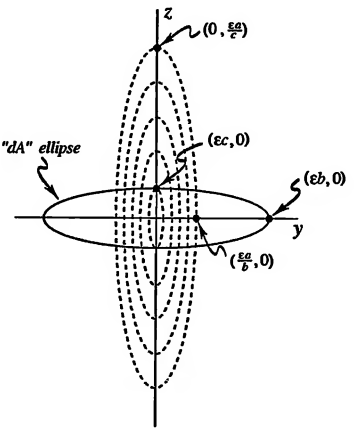


FIGURE 13.9. Near the vertex $(a, 0, 0)$, the five increasingly-large elliptical areas are what's painted by the unit normals to the ellipsoid

$$\frac{x^2}{a^2} + \frac{y^2}{4^2} + \frac{z^2}{1^2} = 1,$$

for $a = 1, 2, 3, 4, 5$, respectively. For a given value of ϵ, the "dA" elliptical area always stays the same, having principal axes ϵb, ϵc and area $A = \pi\epsilon^2 bc$. Dashed ellipses have principal axes $\frac{\epsilon a}{b}$, $\frac{\epsilon a}{c}$ and dimensionless area on the unit sphere of

$$\Theta = \pi \frac{\epsilon a}{b} \frac{\epsilon a}{c}.$$

The ratio

$$\frac{\Theta}{A} \quad \text{simplifies to} \quad \frac{a^2}{b^2 c^2},$$

as in (13.7). Of course an analogous thing happens at each of the other vertices.

Surface Curvature: Some Further Intuition

Student. Can you get the curvature at *other* points of an ellipsoid, too? Maybe even, as with curves, you'll say one can do even better and get a curvature formula working at any point of any smooth surface! [See Appendix 2, p. 387.]

Teacher. In both cases the answer is yes. One can get a curvature formula valid at all points of an ellipsoid (or, more generally, a nondegenerate quadric surface), and it turns out that getting a formula for an arbitrary smooth surface in $\mathbb{R}^3$ is mostly just a glorified version of that.

Let me begin with a couple of basic facts.

The first is this: in analogy to curvature of a curve being measured by a circle of best fit, surface curvature can be measured by the quadric surface of best fit. Suppose that at a point, two different curves are parametrized by power series. If these parametrizations agree up to terms of degree 2, their curvatures are the same at that point.

> Curvature is essentially a "second order" concept—curvature formulas involve only first and second order derivatives.

Student. I need to see an example of this!

Teacher. Take the graph of $y = e^x$, for example. A power series expansion of $y = e^x$ is

$$y = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots.$$

The terms up to degree 2 are

$$y = 1 + x + \frac{x^2}{2!}$$

and define what one could call the "tangent parabola." This and the graph of $y = e^x$ are of course entirely different curves, but they do agree to the second order near $x = 0$. I'm saying that the curvature of the exponential curve at (0, 1) is the same as the curvature of the tangent parabola there.

Student. That seems hard to believe.

Teacher. Just look at our formula,

$$\kappa = \frac{\dot{x}\ddot{y} - \ddot{x}\dot{y}}{((\dot{x})^2 + (\dot{y})^2)^{3/2}},$$

appearing in (13.2) [p. 242]. The highest order of derivative is 2. At the point in question, all derivatives are the same for both curves, so the curvatures agree, too.

Student. Can you work out this example?

Teacher. Sure. By (13.3), the graph of $y = f(x)$ has curvature

$$\kappa = \frac{\ddot{y}}{(1+(\dot{y})^2)^{3/2}},$$

which for our exponential function becomes

$$\kappa = \frac{e^x}{(1+e^{2x})^{3/2}}.$$

So at the point (0, 1), the curvature of the graph is $2^{-3/2}$. For the tangent parabola

$$y = 1 + x + \frac{x^2}{2!},$$

$\dot{y}$ is $1 + x$ which is 1 at $x = 0$; and $\ddot{y} = 1$. So at (0, 1), the parabola's curvature is

$$\kappa = \frac{1}{(1+1)^{3/2}} = 2^{-3/2},$$

which is the same!

Student. That seems profound.

Teacher. And very useful, too, because an analogous thing holds for surfaces. Suppose, for example, that S is the locus of $p(x, y, z) = 0$ for some polynomial p of high degree, and that S^* is the locus for the polynomial p^* consisting of only the terms up through degree 2. Then in the same way that the tangent parabola approximates the exponential graph, the quadric surface defined by $p^* = 0$ approximates S at P, and it does this well enough so that if S is smooth, we can use S^* to compute the curvature at P.

Philosopher. Sounds like progress! So finding the curvature at a point of any smooth surface in $\mathbb{R}^3$ comes down to finding the curvature of a quadric surface—for example, an ellipsoid—at an arbitrary point.

Teacher. Yes. Suppose that at any fixed point P of the quadric, we translate the tangent plane a tiny bit in such a way that it intersects the quadric surface in a curve, exemplified in Figure 13.10. It's an easy exercise to show that the curve must be a conic. Now if P happens to be a vertex point of the ellipsoid, then the normal through P intersects the plane of the conic exactly at its center of symmetry. If, on the other hand, P is *not* a vertex point, then this intersection point is close, but not necessarily *at*, the center of symmetry. However, it turns out that as the amount of translation approaches zero, the intersection point becomes relatively more centered. Geometrically, the situation at a non-extreme point in this way approaches the situation of an *extreme-point* case of the ellipsoid. As the amount of translation approaches zero, our earlier geometric arguments apply ever more exactly.

For example, the limit of the orthogonal principal axes in the moving tangent plane approaches two orthogonal lines. Each such line, together with the normal $\mathbf{n}$ to S at P, defines a plane, and each of these two planes intersects S in a curve having the greatest or least curvature among curves on S cut out by planes containing $\mathbf{n}$. Actually, this is a special case of a more general theorem proved by Euler in 1760. Essentially, it says this:

A theorem of Euler (1760). At a point P of a smooth surface S in $\mathbb{R}^3$, let $\mathbf{v}$ be a vector in the tangent plane to the surface at P, and let N be the plane containing $\mathbf{v}$ and the normal

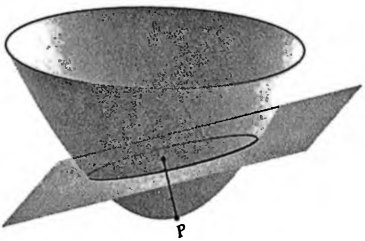


FIGURE 13.10. The plane shown here is a small translate of the tangent plane at an arbitrary point P of an ellipsoid. Although the intersection is an ellipse, the normal through P will not in general intersect the plane at the center of the ellipse. However, as this plane is translated ever closer to P, the normal's intersection gets relatively closer to the center of the shrinking ellipses. If we zoom in to offset the ellipses' diminishing size, we would see the intersection point approach the ellipse center.

to S at P. Let $\kappa(\mathbf{v})$ denote the curvature at P of $S \cap N$. Then there are two orthogonal directions given by vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ along which $\kappa(\mathbf{v})$ attains a maximum value κ_1 and minimum value κ_2. If $\kappa(\mathbf{v})$ is not constant, then the lines containing $\mathbf{v}_1$ and $\mathbf{v}_2$ are uniquely determined. In fact, there's a general formula for $\kappa(\mathbf{v})$: if $\mathbf{v}$ makes an angle α with $\mathbf{v}_1$, then

$$\kappa(\mathbf{v}) = \kappa_1 \cos^2\alpha + \kappa_2 \sin^2\alpha.$$

For example, near (0, 0, 0) the saddle surface $z = x^2 - y^2$, which looks like the saddle on a horse, curves upward fastest along the direction of the horse's length. Its downward curving is greatest in the orthogonal direction where your legs go.

Student. But that theorem seems hard to believe, because a surface could curve upward or downward in directions that *aren't* perpendicular to each other, couldn't it? I once saw an illustration of a "monkey saddle" that had a third region where the surface dipped down for the monkey's tail. It seems there would be a minimum along each of the three dipping-down directions.

Teacher. You raise a good question. I've seen a derivation of an equation for such a monkey saddle. To get three "dips," just use the real or imaginary part of $(x + iy)$ raised to the third power. For example, the real part of $(x + iy)^3$ gives an equation $z = x^3 - 3xy^2$ for the monkey saddle. If you wanted n dips, you could take the real or imaginary part of $(x + iy)^n$.

Student. So that's even more reason to doubt Euler's theorem!

Teacher. But because for any of these surfaces we have an equation, we can find equations for the "$S \cap N$-curves," too. For the monkey saddle, if P is the origin and $z = 0$ is the tangent plane, then $y = mx$ is the equation of a normal plane. Substituting this equation

into $z = x^3 - 3xy^2$ gives $z = (1-m)x^3$. That's a *cubic* with no square term, so according to our formula in (13.2) [p. 242], the curvature at the origin is zero! In fact the surface has zero curvature at P in every v-direction referred to in Euler's theorem. So the theorem continues to hold there.

Student. Even when there are many, many dips?

Teacher. Well, look at the real part of $(x + iy)^n$. Every term has degree n, so if $n > 2$, (13.2) tells us that all those $S \cap N$-curves have zero curvature at the origin. If in the polynomial equation $z = p(x, y)$, every term of p has degree greater than 2, then Euler's theorem holds trivially in the way we just saw.

Philosopher. Let me see if I have the big picture here. We've agreed that finding the curvature at a point of any smooth surface in $\mathbb{R}^3$ comes down to finding the curvature of a quadric surface. And Figure 13.10 suggests that any point of that quadric surface behaves, in the limit, like a vertex point. And I noticed that at a vertex point, the curvature is the product of the coordinate-ellipse curvatures, which themselves are the two extreme curvatures in Euler's theorem.

Teacher. Yes. This means that we need to find the product of the extreme curvatures in Euler's theorem. That's our strategy. That's also now our mission!

Some Algebraic Facts

Teacher. The formula

$$\kappa = \frac{\dot{x}\ddot{y} - \ddot{x}\dot{y}}{((\dot{x})^2 + (\dot{y})^2)^{3/2}}$$

for the curvature at any point of a smooth, parametrized curve looks fancier than

$$\frac{a}{b^2} \quad \text{or} \quad \frac{b}{a^2},$$

the curvatures at ellipse vertex points. And a surface certainly has more variability, or degrees of freedom, than a curve does! So in pursuing our mission, we should be aware that any ultimate formula for surface curvature somehow has to accommodate a greater range of input. This entails more algebra, and to help simplify it, we now introduce some convenient notation.

To begin, we assume that the surface $S \subset \mathbb{R}^3$ is smooth. [See Appendix 2.] This means that some neighborhood of a typical point P in S can be parametrized by

$$\begin{aligned} x &= x(u, v), \\ y &= y(u, v), \\ z &= z(u, v), \end{aligned} \tag{13.10}$$

where x, y, z are smooth functions in u and v throughout that neighborhood. Our first bit of useful notation is to write the parametrization $(x(u, v), y(u, v), z(u, v))$ more compactly as a vector-valued function $\mathbf{r}$ or $\mathbf{r}(u, v)$. This efficiency extends to other basic operations, like differentiation and dot product. For example, $\mathbf{r}_u$ denotes componentwise differentiation

with respect to u. Geometrically, $\mathbf{r}_u$ is a vector tangent to S in the direction of a curve on S along which only u varies. Analogously for $\mathbf{r}_v$. It can be shown that because $S \subset \mathbb{R}^3$ is smooth, $\mathbf{r}_u \times \mathbf{r}_v$ is not the zero-vector; therefore this cross product is normal to S at Q. The normal vector

$$\frac{\mathbf{r}_u \times \mathbf{r}_v}{\sqrt{(\mathbf{r}_u \times \mathbf{r}_v) \cdot (\mathbf{r}_u \times \mathbf{r}_v)}}$$

then has unit length; we denote it by $\mathbf{n}(Q)$, or simply $\mathbf{n}$.

The Leibnitz product rule holds, too: for differentiable, vector-valued functions $\mathbf{f}$ and $\mathbf{g}$ we have

$$(\mathbf{f} \cdot \mathbf{g})' = \mathbf{f}' \cdot \mathbf{g} + \mathbf{f} \cdot \mathbf{g}',$$

where "$'$" denotes differentiation in an arbitrary but fixed direction, and where "$\cdot$" denotes dot product. The rule follows easily from linearity and the Leibnitz rule for the (scalar) components of $\mathbf{f}$ and $\mathbf{g}$:

$$\begin{aligned}
(\mathbf{f} \cdot \mathbf{g})' &= ((f_1, f_2, f_3) \cdot (g_1, g_2, g_3))' \\
&= (f_1 g_1 + f_2 g_2 + f_3 g_3)' \\
&= f_1' g_1 + f_1 g_1' + f_2' g_2 + f_2 g_2' + f_3' g_3 + f_3 g_3' \\
&= f_1' g_1 + f_2' g_2 + f_3' g_3 + f_1 g_1' + f_2 g_2' + f_3 g_3' \\
&= \mathbf{f}' \cdot \mathbf{g} + \mathbf{f} \cdot \mathbf{g}'.
\end{aligned}$$

It often happens that such vector-valued functions $\mathbf{f}$ and $\mathbf{g}$ are orthogonal throughout a neighborhood of P on S. A typical case is when one vector is normal to S and the other lies in the tangent plane. The Leibnitz rule applied to $\mathbf{f} \cdot \mathbf{g} = 0$ means that $\mathbf{f}' \cdot \mathbf{g} = -\mathbf{f} \cdot \mathbf{g}'$. If differentiation is with respect to u then $\mathbf{f}_u \cdot \mathbf{g} = -\mathbf{f} \cdot \mathbf{g}_u$; similarly, $\mathbf{f}_v \cdot \mathbf{g} = -\mathbf{f} \cdot \mathbf{g}_v$. More generally, $\mathbf{f}_s \cdot \mathbf{g} = -\mathbf{f} \cdot \mathbf{g}_s$, where s is arc length along some smooth curve on S. Because $\mathbf{r}_u \cdot \mathbf{n} = \mathbf{r}_v \cdot \mathbf{n} = \mathbf{r}_s \cdot \mathbf{n} = 0$, this gives us a number of "subscript-mover" facts that will prove useful:

$$\begin{aligned}
&\mathbf{r}_{uu} \cdot \mathbf{n} = -\mathbf{r}_u \cdot \mathbf{n}_u; \\
&\mathbf{r}_{uv} \cdot \mathbf{n} = \mathbf{r}_{vu} \cdot \mathbf{n} = -\mathbf{r}_u \cdot \mathbf{n}_v = -\mathbf{r}_v \cdot \mathbf{n}_u; \\
&\mathbf{r}_{vv} \cdot \mathbf{n} = -\mathbf{r}_v \cdot \mathbf{n}_v; \\
&\mathbf{r}_{ss} \cdot \mathbf{n} = -\mathbf{r}_s \cdot \mathbf{n}_s.
\end{aligned} \tag{13.11}$$

A Formula for Surface Curvature

Teacher. We're now ready to begin our journey toward a formula for surface curvature.

Mission: Find the curvature at a point P of a smooth surface $S \subset \mathbb{R}^3$ in terms of the parametrization $\mathbf{r}(u, v)$ of S.

How we'll accomplish the mission: Let T be the tangent plane to S at P. Relative to appropriate coordinates in T, let μ be the slope of a line in T through P. Now this line and the normal $\mathbf{n}(P)$ determine a plane $\mathbf{P}_\mu$ containing them, and $\mathbf{P}_\mu$ in turn intersects S in

a curve C_μ. We will find this curve's curvature κ_μ at P in terms of $\mathbf{r}(u, v)$. From Euler's theorem, we know that if κ_μ isn't constant, then there is exactly one direction along which κ_μ is maximum, and along the perpendicular direction in the tangent plane, κ_μ is minimum. One can arrange that no coordinate axis lie in either of these two extreme directions, and then as $\mathbf{P}_\mu$ spins about $\mathbf{n}(P)$ the extreme directions are determined algebraically by

$$\frac{d\kappa}{d\mu} = 0.$$

This condition leads to a quadratic equation $\kappa^2 + a\kappa + b = 0$ whose roots are the extreme curvature values. The constant term, being the product of these roots, is the curvature of the surface at P. The curvature can therefore be computed in terms of $\mathbf{r}(u, v)$.

Let's begin. Assume C_μ is parametrized by its arc length s, with P corresponding to $s = 0$. For convenience, we write a point of C_μ as $\mathbf{r}(s)$. At a particular point $\mathbf{r}(s) \in C_\mu$, $\mathbf{r}_s(s)$ is the unit tangent vector to C_μ in the direction of increasing s, and $\mathbf{r}_{ss}(s)$ is normal to C_μ. From the definition of a curve's curvature as the amount of turning per unit arc length, a little thought shows that the curvature of C_μ at P is $\kappa_\mu(P) = \mathbf{r}_{ss}(0) \cdot \mathbf{n}(P)$. Now tangent and normal vectors are orthogonal—that is, $\mathbf{r}_s \cdot \mathbf{n} = 0$—and differentiating this gives $\kappa_\mu(P) = -\mathbf{r}_s(0) \cdot \mathbf{n}_s(P)$. We can write this more compactly as $\kappa_\mu = -\mathbf{r}_s \cdot \mathbf{n}_s$. Now there's a good bit of freedom in manipulating differentials:

$$\begin{aligned}
\kappa_\mu = -\mathbf{r}_s \cdot \mathbf{n}_s &= -\frac{d\mathbf{r}}{ds} \cdot \frac{d\mathbf{n}}{ds} = -\frac{d\mathbf{r} \cdot d\mathbf{n}}{d\mathbf{r} \cdot d\mathbf{r}} \\
&= -\frac{(\mathbf{r}_u\, du + \mathbf{r}_v\, dv) \cdot (\mathbf{n}_u\, du + \mathbf{n}_v\, dv)}{(\mathbf{r}_u\, du + \mathbf{r}_v\, dv) \cdot (\mathbf{r}_u\, du + \mathbf{r}_v\, dv)} \\
&= -\frac{(\mathbf{r}_u \cdot \mathbf{n}_u)\, du^2 + (\mathbf{r}_u \cdot \mathbf{n}_v + \mathbf{r}_v \cdot \mathbf{n}_u)\, du\, dv + (\mathbf{r}_v \cdot \mathbf{n}_v)\, dv^2}{(\mathbf{r}_u \cdot \mathbf{r}_u)\, du^2 + (\mathbf{r}_u \cdot \mathbf{r}_v + \mathbf{r}_v \cdot \mathbf{r}_u)\, du\, dv + (\mathbf{r}_v \cdot \mathbf{r}_v)\, dv^2} \\
&= \frac{(\mathbf{r}_{uu} \cdot \mathbf{n})\, du^2 + 2(\mathbf{r}_{uv} \cdot \mathbf{n})\, du\, dv + (\mathbf{r}_{vv} \cdot \mathbf{n})\, dv^2}{(\mathbf{r}_u \cdot \mathbf{r}_u)\, du^2 + 2(\mathbf{r}_u \cdot \mathbf{r}_v)\, du\, dv + (\mathbf{r}_v \cdot \mathbf{r}_v)\, dv^2}. \qquad (13.12)
\end{aligned}$$

The last step comes from further applications of subscript-moving in (13.11). Notice how the minus sign gets absorbed into the numerator in the last line. Also, notice that in (13.12), the differentials' coefficients are all dot products, and therefore scalar quantities. Geometrically, we can think of these coefficients as stretching or shrinking the differentials of length du and dv under the map $\mathbf{r}(u, v)$. It is by means of this stretching and shrinking that flat (u, v)-coordinate patches get morphed into patches on S having curvature. Think of taking a rubber sheet and stretching it smoothly over part of a sphere, ellipsoid or saddle surface. Small rectangular $\Delta u \Delta v$-patches in parameter space generally change size and shape in the process of manufacturing S, and those dot-product scaling factors contain the information about these changes. Of course, since the map $\mathbf{r}(u, v)$ completely determines the surface S, it determines its curvature at every point. Those equations leading to (13.12) show that in fact the dot products coming from $\mathbf{r}(u, v)$ tell us the entire curvature story.

Not surprisingly, since the entire differential theory of surfaces depends on these scaling factors, there is a standard notation for them:

$$E = \mathbf{r}_u \cdot \mathbf{r}_u; \quad F = \mathbf{r}_u \cdot \mathbf{r}_v; \quad G = \mathbf{r}_v \cdot \mathbf{r}_v;$$
$$e = \mathbf{r}_{uu} \cdot \mathbf{n}; \quad f = \mathbf{r}_{uv} \cdot \mathbf{n}; \quad g = \mathbf{r}_{vv} \cdot \mathbf{n}. \tag{13.13}$$

We can therefore write

$$\kappa = \frac{e\,du^2 + 2f\,du\,dv + g\,dv^2}{E\,du^2 + 2F\,du\,dv + G\,dv^2}. \tag{13.14}$$

The denominator is known as the **first fundamental form** of the surface, and the numerator, the **second fundamental form.**

Note that in the first line of the equation sequence leading to (13.12), we've used $(ds)^2 = d\mathbf{r} \cdot d\mathbf{r}$, which appears in the denominator of (13.12) as

$$(\mathbf{r}_u \cdot \mathbf{r}_u)\,du^2 + 2(\mathbf{r}_u \cdot \mathbf{r}_v)\,du\,dv + (\mathbf{r}_v \cdot \mathbf{r}_v)\,dv^2.$$

With our definitions of E, F and G, we see how the first fundamental form relates to the differential of arc length:

$$(ds)^2 = E\,du^2 + 2F\,du\,dv + G\,dv^2.$$

> The differential geometry of a surface depends entirely on its first and second fundamental forms.

Philosopher. What happened to μ? It doesn't seem to appear in (13.14) at all.

Teacher. Differentiation is in a direction tangent to C_μ, so on C_μ, du and dv are not independent: in fact we can interpret the parameter μ as the slope $\frac{dv}{du}$ of the tangent direction. In this sense, (13.14) is general in form; to apply it to a particular C_μ, choose $\frac{dv}{du} = \mu$. Dividing numerator and denominator of (13.14) by du^2 then gives this for the curvature at P in the direction μ:

$$\kappa_\mu = \kappa(\mu) = \frac{e + 2f\mu + g\mu^2}{E + 2F\mu + G\mu^2}. \tag{13.15}$$

To find the extreme values of this curvature, set $\frac{d\kappa}{d\mu} = 0$. After applying the quotient rule, we get

$$\frac{(2f + 2g\mu)(E + 2F\mu + G\mu^2) - (e + 2f\mu + g\mu^2)(2F + 2G\mu)}{(E + 2F\mu + G\mu^2)^2} = 0. \tag{13.16}$$

Of course we want the numerator to be zero—that is,

$$(f + g\mu)(E + 2F\mu + G\mu^2) = (e + 2f\mu + g\mu^2)(F + G\mu). \tag{13.17}$$

If we divide through by the capital-letter factors, we get

$$\frac{e + 2f\mu + g\mu^2}{E + 2F\mu + G\mu^2} = \frac{f + g\mu}{F + G\mu}. \tag{13.18}$$

Now (13.18) was derived from setting $\frac{d\kappa}{d\mu} = 0$, and with this assumption, the left side of (13.18), which is the $\kappa(\mu)$ in (13.15), then simplifies to the right side of (13.18).

Philosopher. This is interesting—in the extreme directions, the curvature κ assumes a simpler form, just as (13.1) [p. 239] is simpler than (13.2) [p. 242].

Teacher. Perhaps you are teaching me to think a little more about symmetry. We chose to look at the slope $\frac{dv}{du} = \mu$, which parametrizes the curve-directions through P, but we could just as well have chosen to look at $\frac{du}{dv}$. Writing $\frac{du}{dv} = \nu$ and mimicking (13.15)–(13.17) leads to this analogue of (13.17):

$$(e\nu + f)(E\nu^2 + 2F\nu + G) = (e\nu^2 + 2f\nu + g)(E\nu + F). \tag{13.19}$$

Substituting $\nu = \frac{1}{\mu}$ into this equation and simplifying gives this mate to (13.18):

$$\frac{e + 2f\mu + g\mu^2}{E + 2F\mu + G\mu^2} = \frac{e + f\mu}{E + F\mu}. \tag{13.20}$$

We therefore have *two* simple ways of writing the extreme curvatures:

$$\kappa = \frac{f + g\mu}{F + G\mu}; \qquad \kappa = \frac{e + f\mu}{E + F\mu}. \tag{13.21}$$

After clearing denominators, it is easy to solve for μ in terms of κ, getting these counterparts to (13.21):

$$\mu = \frac{e - E\kappa}{F\kappa - f}; \qquad \mu = \frac{f - F\kappa}{G\kappa - g}. \tag{13.22}$$

By equating right-hand sides of the equations in (13.22), we can get an equation for κ in terms of E, F, G, e, f, g:

$$\frac{e - E\kappa}{F\kappa - f} = \frac{f - F\kappa}{G\kappa - g},$$

which can be rewritten as

$$(E\kappa - e)(G\kappa - g) - (F\kappa - f)^2 = 0.$$

Multiplying this out and collecting like powers of κ gives

$$(EG - F^2)\kappa^2 - (Eg + eG - 2Ff)\kappa + (eg - f^2) = 0.$$

This is quadratic in κ. Dividing through by $EG - F^2$ makes it monic, with constant term

$$\frac{eg - f^2}{EG - F^2}$$

and, of course, this constant term is the product of its two roots—the two extreme values of κ. This constant term is therefore the promised surface curvature (!):

$$\kappa = \frac{eg - f^2}{EG - F^2}. \tag{13.23}$$

The single scalar quantity κ provides a lot of information about how a surface bends. But in our real world of bending 4-D space-time, no single scalar quantity can adequately capture the many subtleties of curvature. Einstein used "Riemann curvature tensors," essentially symmetric 4×4 matrices having $10 (= 4 + 3 + 2 + 1)$ different entries.

Philosopher. That is remarkable for its simple form. But one thing bothers me. In our basic strategy, "How we'll accomplish the mission," we seem to have assumed that the curvature is not constant. We assumed this in using Euler's theorem to get well-defined perpendicular directions for maximum and minimum curvatures. Does our elegant formula

$$\kappa = \frac{eg - f^2}{EG - F^2}$$

still work, say, on a sphere, where the curvature is constant?

Philosopher. Let's see! Take a sphere of radius R. We know it has curvature $\frac{1}{R^2}$ at every point. Working out E, F, G, e, f, g and substituting these into

$$\frac{eg - f^2}{EG - F^2}$$

should give $\frac{1}{R^2}$. Now one parametrization of a sphere comes directly from spherical coordinates:

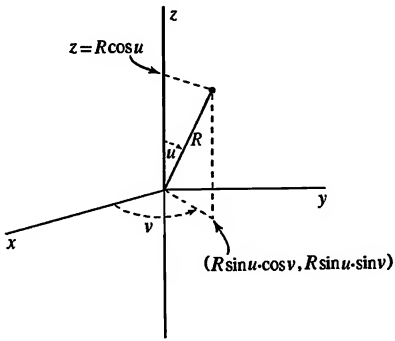


FIGURE 13.11. Here, Eulerian angles are denoted by u and v, and we can use them as the u and v in (13.10) on p. 254. The z-coordinate of a typical point on the sphere is $R \cos u$, and the dashed triangle base in the (x, y)-plane has length $R \sin u$. The x- and y-components of the typical point are therefore $R \sin u \cos v$ and $R \sin u \sin v$.

Figure 13.11 leads to this parametrization $\mathbf{r}$ for a sphere of radius R:

$$\mathbf{r} = (R\sin u \cdot \cos v,\ R\sin u \cdot \sin v,\ R\cos u).$$

Computing the various derivatives of $\mathbf{r}$ is straightforward:

$$\mathbf{r}_u = (R\cos u \cdot \cos v,\ R\cos u \cdot \sin v,\ -R\sin u);$$
$$\mathbf{r}_v = (-R\sin u \cdot \sin v,\ R\sin u \cdot \cos v,\ 0);$$
$$\mathbf{r}_{uu} = (-R\sin u \cdot \cos v,\ -R\sin u \cdot \sin v,\ -R\cos u);$$
$$\mathbf{r}_{uv} = (-R\cos u \cdot \sin v,\ R\cos u \cdot \cos v,\ 0);$$
$$\mathbf{r}_{vv} = (-R\sin u \cdot \cos v,\ -R\sin u \cdot \sin v,\ 0);$$
$$n = (\sin u \cdot \cos v,\ \sin u \cdot \sin v,\ \cos u).$$

Writing the basic dot products for E, F, G, e, f, g gives, after a little simplification:

$$E = \mathbf{r}_u \cdot \mathbf{r}_u = R^2(\cos^2 u \cdot \cos^2 v + \cos^2 u \cdot \sin^2 v + \sin^2 u) = R^2;$$
$$F = \mathbf{r}_u \cdot \mathbf{r}_v = R^2(-\cos u \cdot \cos v \cdot \sin u \cdot \sin v + \cos u \cdot \cos v \cdot \sin u \cdot \sin v) = 0;$$
$$G = \mathbf{r}_v \cdot \mathbf{r}_v = R^2(\sin^2 u \cdot \sin^2 v + \sin^2 u \cdot \cos^2 v) = R^2\sin^2 u;$$
$$e = \mathbf{r}_{uu} \cdot \mathbf{n} = -R(\sin^2 u \cdot \cos^2 v + \sin^2 u \cdot \sin^2 v + \cos^2 u) = -R;$$
$$f = \mathbf{r}_{uv} \cdot \mathbf{n} = -R(-\cos u \cdot \sin v \cdot \sin u \cdot \cos v + \cos u \cdot \cos v \cdot \sin u \cdot \sin v) = 0;$$
$$g = \mathbf{r}_{vv} \cdot \mathbf{n} = -R(\sin^2 u \cdot \cos^2 v + \sin^2 u \cdot \sin^2 v) = -R\sin^2 u.$$

Our final computation is then

$$\kappa = \frac{eg - f^2}{EG - F^2} = \frac{(-R)\cdot(-R\sin^2 u)}{R^4\sin^2 u} = \frac{1}{R^2}.$$

Student. That *does* work out nicely!

Einstein's general theory of relativity tells us that mass "bends" space-time, somewhat the way a cannonball creates a depression in a trampoline's surface. A small marble, if it's moving too slowly, won't have enough energy to break free from the influence of the curvature, and will stay trapped in an orbit around the cannonball. Of course there can be more than one marble—perhaps also, say, a golf ball, a steel ball bearing and so on, all going in different orbits. This is basically what all the various planets are doing as they orbit our sun.

A Few Reality Checks

Student. I have a question. Can you do for a more general, parametrized surface in $\mathbb{R}^3$ what you did before for a general, parametrized curve in $\mathbb{R}^2$? That curve was parametrized by $x = x(t)$, $y = y(t)$, and we found a concrete expression for the curvature in terms of these functions—that is,

$$\kappa = \frac{\dot{x}\ddot{y} - \ddot{x}\dot{y}}{((\dot{x})^2 + (\dot{y})^2)^{3/2}}.$$

Teacher. You're asking if there's a concrete expression for surface curvature in terms of the surface's parametrization $x = x(u, v)$, $y = y(u, v)$, $z = z(u, v)$, as in (13.10)?

Student. Exactly.

Teacher. That means translating each of E, F, G, e, f, g into x, y, z. Let's start with E, F, G.

$$\begin{aligned} E &= \mathbf{r}_u \cdot \mathbf{r}_u = x_u^2 + y_u^2 + z_u^2; \\ F &= \mathbf{r}_u \cdot \mathbf{r}_v = x_u x_v + y_u y_v + z_u z_v; \\ G &= \mathbf{r}_v \cdot \mathbf{r}_v = x_v^2 + y_v^2 + z_v^2. \end{aligned} \tag{13.24}$$

Of course $EG - F^2$ involves quite a few terms—all involving x, y, z—but some of these terms cancel out. It turns out that the remaining ones fall into a sum of squares:

$$EG - F^2 = (y_u z_v - y_v z_u)^2 + (x_u z_v - x_v z_u)^2 + (x_u y_v - x_v y_u)^2. \tag{13.25}$$

Student. So far, not too bad.

> When the blind beetle crawls over the surface of a globe, he doesn't realize that the track he has covered is curved. I was lucky enough to have spotted it.
>
> —Albert Einstein

Teacher. Now for e, f and g. Because each of these is a dot product with the unit normal $\mathbf{n}$, let's first find $\mathbf{n}$. We noted earlier that

$$\mathbf{n} = \frac{\mathbf{r}_u \times \mathbf{r}_v}{\sqrt{(\mathbf{r}_u \times \mathbf{r}_v) \cdot (\mathbf{r}_u \times \mathbf{r}_v)}}. \tag{13.26}$$

Now

$$\mathbf{r}_u \times \mathbf{r}_v = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ x_u & y_u & z_u \\ x_v & y_v & z_v \end{vmatrix} = (y_u z_v - y_v z_u)\mathbf{i} - (x_u z_v - x_v z_u)\mathbf{j} + (x_u y_v - x_v y_u)\mathbf{k},$$

so the denominator of (13.26) is

$$\sqrt{(y_u z_v - y_v z_u)^2 + (x_u z_v - x_v z_u)^2 + (x_u y_v - x_v y_u)^2}. \tag{13.27}$$

Student. This is just the square root of $EG - F^2$ in (13.25)!

Teacher. Another little miracle, isn't it? Now from (13.13) and (13.26), we see that the denominator of $eg - f^2$ is (13.27), and that its numerator is

$$[\mathbf{r}_{uu} \cdot (\mathbf{r}_u \times \mathbf{r}_v)] \cdot [\mathbf{r}_{vv} \cdot (\mathbf{r}_u \times \mathbf{r}_v)] - [\mathbf{r}_{uv} \cdot (\mathbf{r}_u \times \mathbf{r}_v)]^2. \tag{13.28}$$

In forming the dot products in (13.28), any cofactor of the determinant

$$\mathbf{r}_u \times \mathbf{r}_v = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ x_u & y_u & z_u \\ x_v & y_v & z_v \end{vmatrix}$$

is multiplied by the corresponding component of $\mathbf{r}_{uu}$, $\mathbf{r}_{uv}$ or $\mathbf{r}_{vv}$, so each of these dot products is a 3×3 determinant of scalar functions. Putting everything together gives, for $\frac{eg-f^2}{EG-F^2}$,

$$\kappa = \frac{\begin{vmatrix} x_{uu} & y_{uu} & z_{uu} \\ x_u & y_u & z_u \\ x_v & y_v & z_v \end{vmatrix} \begin{vmatrix} x_{vv} & y_{vv} & z_{vv} \\ x_u & y_u & z_u \\ x_v & y_v & z_v \end{vmatrix} - \begin{vmatrix} x_{uv} & y_{uv} & z_{uv} \\ x_u & y_u & z_u \\ x_v & y_v & z_v \end{vmatrix}^2}{[(y_u z_v - y_v z_u)^2 + (x_u z_v - x_v z_u)^2 + (x_u y_v - x_v y_u)^2]^2}. \tag{13.29}$$

Student. Well, I asked, and you delivered! So (13.29) is the analogue of

$$\kappa = \frac{\dot{x}\ddot{y} - \ddot{x}\dot{y}}{((\dot{x})^2 + (\dot{y})^2)^{3/2}},$$

eh?

> Of the general theory of relativity, it's been said that "mass tells space-time how to warp, and the warping of space-time tells mass how to move."

Teacher. Yes.

Student. Does (13.29) check out dimensionally?

Teacher. Good question. We should get an overall dimension of L^{-2}. Now $\mathbf{r}(u, v)$ maps from the (u, v)-plane into $\mathbb{R}^3$, and at any point of the (u, v)-plane, differentials du and dv determine a differential ds in $\mathbb{R}^3$ via $ds^2 = E\,du^2 + 2F du dv + G\,dv^2$. The dimension of ds is L, and that of products like du^2 or $du\,dv$ is L^2. For quotients, work by analogy: if s represents distance and t, time, then the quotient $\frac{ds}{dt}$, which represents velocity, has dimension LT^{-1}. And $\frac{d^2s}{dt^2}$, representing acceleration—distance per time per time—has dimension LT^{-2}. Dimensions of derivatives in (13.29) are determined similarly: x_u, for example, has dimension LL^{-1}—it's dimensionless. And x_{uu} has dimension $LL^{-2} = L^{-1}$. The numerator in (13.29) thus has dimension L^{-2}, the denominator is dimensionless, so the overall dimension is L^{-2}.

Philosopher. Each of E, F, and G is dimensionless, but e, f, and g each has dimension L^{-1}, right?

Teacher. Yes. Notice that we normalize $\mathbf{n}$ by dividing by its length, leaving it dimensionless.

Philosopher. That fits in with $\kappa_\mu = -\mathbf{r}_s \cdot \mathbf{n}_s$ at the beginning of the equation sequence ending with (13.12): $\mathbf{r}_s$ is dimensionless and $\mathbf{n}_s$ has dimension L^{-1}, so κ_μ has dimension L^{-1}, as it should.

> Newton felt uncomfortable with the idea of "action at a distance." How can the sun somehow manage to sense the presence of a planet, and then pull on it? The answer is that the sun alters the curvature of space-time. Because of gravity's inverse-square strength, its influence spreads outward to the entire universe. (Einstein's general theory implies that this spreading takes place at the speed of light, but accurately measuring this tiny effect has proved immensely challenging.) Again because of the inverse square law, space is more intensely curved near a large mass like the sun or a planet, and we, along with other "test particles," feel this curvature as a gravitational force field.

Theorema Egregium

Philosopher. We have an explicit formula (13.29) for the curvature of a parametrized surface in $\mathbb{R}^3$. But what about the curvature of a complex circle? Or of a more general complex conic? Those are surfaces, but they live in $\mathbb{R}^4$, not $\mathbb{R}^3$. It seems that this could be a major problem.

Teacher. You raise an important issue. Although (13.29) is an explicit formula for

$$\frac{eg - f^2}{EG - F^2},$$

we see from (13.13) [p. 257] that each of e, f, and g is defined in terms of a normal to the surface. This means that information about how the surface S is positioned in $\mathbb{R}^3$ is used in computing e, f, and g—and therefore also in computing κ. But, for example, a surface in $\mathbb{R}^4$ doesn't have a unique normal line at each point. Instead, through any point P of such a smooth surface, there's a whole *plane* orthogonal to S at P, so there's a problem in getting an analogue of $\mathbf{n}$ in (13.13).

Philosopher. What about E, F, and G? Is there a problem with them, too?

Teacher. No. These are metric stretch-and-shape data, and are intrinsic to the surface. An intelligent "microbe" carrying out a geodetic survey on a smooth surface, oblivious to any surrounding space the surface might lie in, could gather enough information to determine E, F, and G. For all it knows—or cares—the surface might live in 3, 4, or 100 dimensions. None of that would make any difference. Determining e, f, and g by using the definitions (13.13) is a different story, because a normal vector appears in each of them. But if we

could somehow carry out the determination of the *ratio*

$$\frac{eg - f^2}{EG - F^2}$$

using strictly in-surface measurements, with no reference to anything outside, we could then use

$$\frac{eg - f^2}{EG - F^2}$$

to actually *define* curvature on surfaces lying in any $\mathbb{R}^n$.

Philosopher. Is that possible?

Teacher. Yes, and the result is the essence of Gauss' famous *Theorema egregium*. In his diary—written in Latin, as he wrote all his mathematics—he calls it "a most excellent theorem." It is often stated as "curvature is a bending invariant," since pure bending doesn't affect in-surface measurements. For example, bending a piece of graph paper doesn't change the distance between rulings or area of its squares.

> There's another side to the coin "curvature is a bending invariant." If you stretch or shrink the material of a surface (rather than merely bend it) you generally introduce changes in curvature. In drawing a terrestrial map on a flat surface, one must do more than bend. It's a fact that you can't map any spherical area to a plane without some distortion, and because of this, there's an array of different approaches to mapping the earth onto a flat piece of paper, each minimizing some aspect of distortion. For example, a "cylindrical map" preserves area, a stereographic map (as in Figure 8.8 or 8.9 on p. 133 or 134) preserves angles, as does the well-known Mercator projection of the earth onto a rectangle.

For such a great theorem, the modern proof is surprisingly short, mainly depending on some clever algebra. Here's the idea:

Let A and B be two 3×3 matrices. (They could just as well be $n \times n$, but we need only 3×3.) Let A_i be the ith row of A, and B_j, the jth column of B. The matrix product

$$AB = \begin{pmatrix} A_1 \\ A_2 \\ A_3 \end{pmatrix} (B_1 \quad B_2 \quad B_3)$$

can be written as

$$\begin{pmatrix} A_1 \cdot B_1 & A_1 \cdot B_2 & A_1 \cdot B_3 \\ A_2 \cdot B_1 & A_2 \cdot B_2 & A_2 \cdot B_3 \\ A_3 \cdot B_1 & A_3 \cdot B_2 & A_3 \cdot B_3 \end{pmatrix}. \tag{13.30}$$

Now the problem with (13.29) lies in its numerator. However, the determinant of a matrix is the determinant of its transpose, so this numerator can be written as

$$\underbrace{\begin{vmatrix} x_{uu} & y_{uu} & z_{uu} \\ x_u & y_u & z_u \\ x_v & y_v & z_v \end{vmatrix}}_{A} \underbrace{\begin{vmatrix} x_{vv} & x_u & x_v \\ y_{vv} & y_u & y_v \\ z_{vv} & z_u & z_v \end{vmatrix}}_{B} - \underbrace{\begin{vmatrix} x_{uv} & y_{uv} & z_{uv} \\ x_u & y_u & z_u \\ x_v & y_v & z_v \end{vmatrix}}_{C} \underbrace{\begin{vmatrix} x_{uv} & x_u & x_v \\ y_{uv} & y_u & y_v \\ z_{uv} & z_u & z_v \end{vmatrix}}_{D} \tag{13.31}$$

We can now apply (13.30) to matrices A and B, and also to C and D—that is, to

$$A = \begin{pmatrix} A_1 \\ A_2 \\ A_3 \end{pmatrix} = \begin{pmatrix} \mathbf{r}_{uu} \\ \mathbf{r}_u \\ \mathbf{r}_v \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} B_1 & B_2 & B_3 \end{pmatrix} = \begin{pmatrix} \mathbf{r}'_{vv} & \mathbf{r}'_u & \mathbf{r}'_v \end{pmatrix},$$

with analogous expressions for C and D. Then (13.31) looks like this:

$$\begin{vmatrix} \mathbf{r}_{uu}\cdot\mathbf{r}_{vv} & \mathbf{r}_{uu}\cdot\mathbf{r}_u & \mathbf{r}_{uu}\cdot\mathbf{r}_v \\ \mathbf{r}_u\cdot\mathbf{r}_{vv} & \mathbf{r}_u\cdot\mathbf{r}_u & \mathbf{r}_u\cdot\mathbf{r}_v \\ \mathbf{r}_v\cdot\mathbf{r}_{vv} & \mathbf{r}_v\cdot\mathbf{r}_u & \mathbf{r}_v\cdot\mathbf{r}_v \end{vmatrix} - \begin{vmatrix} \mathbf{r}_{uv}\cdot\mathbf{r}_{uv} & \mathbf{r}_{uv}\cdot\mathbf{r}_u & \mathbf{r}_{uv}\cdot\mathbf{r}_v \\ \mathbf{r}_u\cdot\mathbf{r}_{uv} & \mathbf{r}_u\cdot\mathbf{r}_u & \mathbf{r}_u\cdot\mathbf{r}_v \\ \mathbf{r}_v\cdot\mathbf{r}_{uv} & \mathbf{r}_v\cdot\mathbf{r}_u & \mathbf{r}_v\cdot\mathbf{r}_v \end{vmatrix}, \tag{13.32}$$

so from (13.29), the surface curvature is (13.32) divided by

$$((y_u z_v - y_v y_u)^2 + (x_u z_v - x_v z_u)^2 + (x_u y_v - x_v y_u)^2)^2 = (EG - F^2)^2. \tag{13.33}$$

The size of each determinant is 3×3, reflecting the 2-dimensionality of the surface. Each vector $\mathbf{r}$ (and its derivatives) has three components, reflecting the 3-dimensionality of the surrounding space $\mathbb{R}^3$. But in (13.32), notice that each dot product makes sense, *no matter how many components* $\mathbf{r}$ *has*! If S lies in $\mathbb{R}^4$, then the vector $\mathbf{r}(u, v)$ has four components. These dot products are similar to the ones in (13.24) [p. 261], but with four summands instead of three. Such dot products continue to reflect intrinsic metric properties of S, just as they do when S lives in $\mathbb{R}^3$.

So where are we now? In establishing the *Theorema egregium*, we have in fact carried out the determination of the ratio

$$\frac{eg - f^2}{EG - F^2}$$

using strictly in-surface measurements, with no reference to anything outside. So as we said earlier, we can now use

$$\frac{eg - f^2}{EG - F^2}$$

to *define* curvature on surfaces lying in any $\mathbb{R}^n$!

Student. Just like that, you *define* it?!

Teacher. That is a typical state of affairs when you want to generalize to higher dimensions. For example in one, two or three dimensions, the distance between two points can be defined using the Pythagorean point-to-point formula, and we can simply extend the

definition of distance to four and higher dimensions by using analogous formulas. Same idea for orthogonality, angles, and so on. The *Theorema egregium* gives us an algebraic formula for curvature which we can use, in the same spirit as the point-to-point formula, to define surface curvature when the surface lives in $\mathbb{R}^n$.

Philosopher. So we find a parametrization of a complex circle in $\mathbb{C}^2 = \mathbb{R}^4$, substitute into (13.32), divide by $(EG - F^2)^2$—and we have what we call its *curvature*!

Teacher. That's right.

> Efficient bookkeeping methods such as vectors and matrices greatly aid us in many everyday computations. Gauss made his discoveries with no such modern advantages.

Chapter Summary

This chapter begins with the curvature of real curves, which is intuitively defined as the amount of turning per unit of arc length. This idea is used to find the curvature of an ellipse at its vertices, and is extended to a formula for the curvature at any point of a smooth curve in $\mathbb{R}^2$.

The chapter next turns to the curvature of surfaces in $\mathbb{R}^3$. In defining surface curvature, we work by analogy with curves: a curve's curvature can be regarded as the limiting ratio of angular change $\Delta\theta$ of its normals to a little piece Δs of the curve as $\Delta s \to 0$. Similarly, the curvature of a surface is the limiting ratio of *solid-angle* turning done by the surface normals over a little piece ΔA of the surface as $\Delta A \to 0$. In analogy with the above ellipse computation, we use our intuitive concept to find the surface curvature at the vertices of an ellipsoid. A main accomplishment of the chapter is doing for parametrized surfaces what we did for parametrized curves: find a curvature formula at any point of a smooth (parametrized) surface in $\mathbb{R}^3$. [See Appendix 2.]

Since our complex circle lives in $\mathbb{R}^4$, we need to extend to $\mathbb{R}^4$ our ideas about curvature of surfaces in $\mathbb{R}^3$. We do this by proving Gauss' *Theorema egregium*, which expresses surface curvature strictly in terms of *in-surface* measurements. We then use the main formula of *Theorema egregium* to *define* curvature of two-dimensional surfaces when the surrounding space has arbitrarily high real dimension.

Exercises

1. The real saddle surface $z = x^2 - y^2$ has the obvious parametrization $(x, y, z) = (u, v, u^2 - v^2)$. Use (13.29) on p. 262 and Maple to sketch in (u, v, κ)-space a curvature plot of this surface. Note that the curvature of the saddle surface is everywhere negative, but that its magnitude diminishes rapidly as one moves away from the origin.
2. A "tractrix" is sometimes described as the curve traced out by an old-fashioned watch fob as the free end of its weightless chain is slowly pulled along a straight line (or, equivalently, as the path of a dog chasing a rabbit running along a straight path). Here

are parametric equations for one such curve:

$$x = t - \tanh(t); \quad y = \frac{1}{\cosh(t)}.$$

(a) Have Maple graph this curve. Then find its curvature at each of its points.

(b) Find parametric equations for the surface obtained by revolving this tractrix about the x-axis. Have Maple plot this surface.

(c) The surface in (b) is called a *pseudosphere*. Like a sphere, its curvature is the same at every point where it can be computed (we'll call such points "smooth"), but is negative. Use (13.29) to find that negative curvature.

(d) What points are "not smooth"? Why?

(e) Given any negative number, can you find a pseudosphere having that as its curvature? What's the analogy with an ordinary sphere?

3. Thinking of Gauss curvature as the product of extreme curvatures as in the text, compare the curvature of the pseudosphere with the curvature of the tractrix at x_o times the curvature of the circle in which the plane $x = x_o$ cuts the pseudosphere. Geometrically, why does this approximation to the actual curvature get better as x_o increases?

4. Use Maple to extend Figure 13.6 on p. 243 to a surface plot in (t, a, κ)-space of the curvature of

$$\frac{x^2}{a^2} + y^2 = 1,$$

as a varies from 0.2 to, say, $a = 6$.

5. Let a plane rotate around the z-axis in $\mathbb{R}^3$. Consider in that plane a circle of radius r whose center is at a distance $R > r$ from the z-axis. As the plane spins, the circle sweeps out a torus. Suppose an angle u parametrizes the plane's rotation. Show that the following parametrizes the torus:

$$x(u, v) = (R + r\cos v)\cos u;$$

$$y(u, v) = (R + r\cos v)\sin u;$$

$$z(u, v) = r\sin v.$$

(a) Choosing $R = 3$ and $r = 1$, have Maple sketch the torus' curvature as a surface in (u, v, κ)-space. Note its double periodicity (in u and v).

(b) By looking only at the torus, say which values of u and v correspond to "saddle shape," or negative, curvature. Does your graph bear out your answer?

(c) Do the same as in (b) for points of positive curvature, and also for points of zero curvature.

6. A real circle living in $\mathbb{R}^2$ is very simple and natural, and as inhabitants of $\mathbb{R}^4$ will tell you, so is the set-product of two circles in $\mathbb{R}^4$—sometimes called a "flat torus." The parametrization in Exercise 5 is not especially beautiful, and that's because the nice torus in $\mathbb{R}^4$ had to be curled up and forced into $\mathbb{R}^3$ in order for us 4-D-challenged humans to "see a torus"!

(a) Show that a parametrization of a torus in $\mathbb{R}^4$ is

$$(x_1, x_2, x_3, x_4) = (\cos u, \sin u, \cos v, \sin v).$$

(b) Show that this torus lies on a sphere S^3 in $\mathbb{R}^4$.

(c) Plot in (u, v, κ)-space the Gauss curvature of this torus. Would you say that the "flat torus" is well-named?

14

Curvature of Conics

in which our trio discovers that the Gauss curvature of complex ellipses and hyperbolas is negative and varies wildly

The Curvature of Complex Circles

Philosopher. Once we have a smooth parametrization of a smooth surface, we can use (13.32) and (13.33) [p. 265] to obtain its Gauss curvature. And we've already found a parametrization of the image in $\mathbb{R}^4$ of the unit complex circle. In terms of u and v, this parametrization in (12.4) [p. 217] becomes

$$\mathbf{r} = \mathbf{r}(u, v) = (\cos u \cosh v, -\sin u \sinh v, \sin u \cosh v, \cos u \sinh v). \tag{14.1}$$

This comes from the complex parametrization

$$x = x_1 + ix_2 = \cos(u + iv); \quad y = y_1 + iy_2 = \sin(u + iv).$$

[See the box on p. 224.]

Teacher. We can introduce stretching-factor coefficients so that for the ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1,$$

(14.1) gives us this parametrization:

$$\mathbf{r}(u, v) = (a \cos u \cosh v, -a \sin u \sinh v, b \sin u \cosh v, b \cos u \sinh v). \tag{14.2}$$

Philosopher. There must be a mate to that for the hyperbola

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1.$$

Teacher. There is. To get it, just mimic what we did for the ellipse, using the hyperbolic analogue

$$x = x_1 + ix_2 = \cosh(u + iv); \quad y = y_1 + iy_2 = \sinh(u + iv).$$

Student. But you need sum-of-angle formulas for those hyperbolic functions!

Teacher. That's right. We've met them before [See the box on p. 224], and here they are:

$$\cosh(u+iv) = \cosh u \cdot \cos v + i \sinh u \cdot \sin v;$$

$$\sinh(u+iv) = \sinh u \cdot \cos v + i \cosh u \cdot \sin v.$$

These lead to parametric equations for the hyperbola

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1,$$

and they turn out to be the same as (14.2), but with circular and hyperbolic functions everywhere interchanged:

$$\mathbf{r}(u, v) = (a \cosh u \cos v, -a \sinh u \sin v, b \sinh u \cos v, b \cosh u \sin v). \tag{14.3}$$

Student. To get the curvature of a complex ellipse or hyperbola, we substitute parametrizations (14.2) or (14.3) into formulas (13.32) and (13.33) [p. 265], right?

Teacher. Yes. In the case of our complex ellipse, for example, we have:

$$\mathbf{r}_u = (-a \sin u \cosh v, -a \cos u \sinh v, b \cos u \cosh v, -b \sin u \sinh v);$$

$$\mathbf{r}_v = (a \cos u \sinh v, -a \sin u \cosh v, b \sin u \sinh v, b \cos u \cosh v);$$

$$\mathbf{r}_{uu} = (-a \cos u \cosh v, a \sin u \sinh v, -b \sin u \cosh v, -b \cos u \sinh v);$$

$$\mathbf{r}_{uv} = (-a \sin u \sinh v, -a \cos u \cosh v, b \cos u \sinh v, -b \sin u \cosh v);$$

$$\mathbf{r}_{vv} = (a \cos u \cosh v, -a \sin u \sinh v, b \sin u \cosh v, b \cos u \sinh v).$$

When $a = b$, the dot products simplify quite a bit. For example, when $a = b = 1$—the unit circle—(13.32) becomes

$$\begin{vmatrix} -\cosh 2v & 0 & -\sinh 2v \\ 0 & \cosh 2v & 0 \\ \sinh 2v & 0 & \cosh 2v \end{vmatrix} - \begin{vmatrix} \cosh 2v & \sinh 2v & 0 \\ \sinh 2v & \cosh 2v & 0 \\ \sinh 2v & 0 & \cosh 2v \end{vmatrix}. \tag{14.4}$$

In simplifying, we've used the hyperbolic double-angle formulas

$$\cosh^2 v + \sinh^2 v = \cosh 2v \quad \text{and} \quad 2 \cdot \sinh v \cdot \cosh v = \sinh 2v.$$

The first determinant of (14.4) is $-\cosh 2v$, and the second one is $\cosh 2v$. So (14.4) itself is $-2\cosh 2v$. Now

$$E = \mathbf{r}_u \cdot \mathbf{r}_u, F = \mathbf{r}_u \cdot \mathbf{r}_v, G = \mathbf{r}_v \cdot \mathbf{r}_v,$$

which in our example turns out to be $E = G = \cosh 2v$, $F = 0$. That makes

$$(EG - F^2)^2 = \cosh^4 2v,$$

so we now have the curvature of the unit circle: it is (14.4) divided by $(EG - F^2)^2$:

$$\kappa_{\text{unit circle}} = \frac{-2}{\cosh^3 2v}. \tag{14.5}$$

Philosopher. A negative sign! I know we observed that Gauss assigned a sign depending on whether the Gauss map happens to be orientation-preserving or orientation-reversing, but for that sign just to fall out algebraically—that seems remarkable.

Teacher. There seems to be evidence that the complex circle is actually some sort of "saddle surface"! That's something to explore.

Student. I see something that looks strange to me—there's no u in that formula!

Teacher. Well, in a parametrization, choosing any value of v defines a "u-curve," along which only u varies, so the formula just says that the curvature is constant on any u-curve. The most famous of these u-curves is the ordinary real unit circle, which comes from setting $v = 0$ in the complex parametrization (14.1). In this case the parametrization becomes $(\cos u, 0, \sin u, 0)$. Of course, writing only the usual real coordinates gives the familiar real parametrization $(\cos u, \sin u)$.

Philosopher. But when $v = 0$, your curvature formula (14.5) for the unit circle gives -2. That seems hard to believe... shouldn't the curvature of the unit circle be 1?

Teacher. The curvature of the *real* unit circle is 1, but the *complex* unit circle is a surface in $\mathbb{R}^4$, and there's more going on. Remember that even for a surface living in the more restricted environment of $\mathbb{R}^3$, as we walk around within a little patch of the surface, the unit normal correspondingly wiggles around, painting an image patch on the unit sphere in $\mathbb{R}^3$—and the size of the curvature is the ratio of the image patch's area to the original patch. Well, in $\mathbb{R}^4$, there's not just a *line* normal to the tangent plane at a point, but rather a whole *plane*. An orthonormal basis in this normal plane thus consists of two unit vectors, each a paint brush! Intuitively, as we walk around a little surface-patch in $\mathbb{R}^4$, each of them wiggles around, painting its own little patch. So the total painted area (the numerator of our ratio) comes from two sources—from each vector in the basis. In $\mathbb{R}^n$, there would be $(n-2)$ wiggling paint brushes. Remarkably, it turns out that *the total painted area is independent of the choice of orthonormal basis.*

Student. A bigger-sized numerator, and the same-sized denominator—that *would* make the result larger-sized!

Philosopher. Suppose you took an ordinary sphere in $\mathbb{R}^3$ and added another dimension, making that same sphere lie in $\mathbb{R}^4$. That certainly shouldn't change the sphere's curvature!

Teacher. I agree. In that case, the original normal to the sphere does its usual wiggle, but nothing is happening in the new dimension, and the normal in that dimension doesn't move. Therefore, no extra painted area is created.

Philosopher. Let me try to get things clear, here. When the surface lives in $\mathbb{R}^3$, we get a limiting ratio of painted image (solid angle) to surface-patch area, and we saw that this ratio factors into a product of two extreme sectional curvatures: we can write, for the size, $|\kappa| = |\kappa_1||\kappa_2|$. If there are *two* paint brushes working away in $\mathbb{R}^4$, does each one contribute such a product?

Teacher. That's the idea.

Student. That idea is as clear as milk to me. I don't feel completely comfortable with either the 2 or the minus sign. It would really help if you could work through the complex circle example with pictures and computations so I can see where things come from.

Teacher. A very good and reasonable request. We should all learn something!

Let's begin by considering one of the paintbrushes—that is, an orthonormal basis vector $\mathbf{n}$ in the normal plane. At a point P of the surface, as patches containing P grow smaller and smaller, the brush's motions become ever tighter. As a way to compute the limiting ratio, we will do something similar to what we did earlier in the $\mathbb{R}^3$ case. As you remember, we formed the plane containing the normal and a unit vector tangent to the surface, and we intersected this plane with the surface, getting a curve. As we swept through all these unit vectors, the curvatures of these curves attained a maximum and minimum, whose product is the limiting ratio—that is, the size of the curvature of the surface at P.

Now suppose the surrounding space is $\mathbb{R}^4$. We similarly form the plane containing a normal and a unit vector $\mathbf{t}$ in the tangent plane $\mathbf{T}$ to the surface. However, in this case this plane will usually intersect the surface in only the point P, not a curve. This happens because we're not in $\mathbb{R}^3$, but in a larger space, $\mathbb{R}^4$—think, for example, of how in real (x_1, x_2, y_1, y_2)-space, the (x_1, x_2)- and (y_1, y_2)-planes intersect not in a line, but only in a point.

In spite of this, there will be a curve on the surface having $\mathbf{t}$ as a tangent vector, and the curve has a *curvature vector*, defined to be $\frac{d\mathbf{t}}{ds}$. In fact, its component in the $\mathbf{n}$ direction happens to depend only on $\mathbf{t}$, not the particular curve on S having $\mathbf{t}$ as tangent.

Philosopher. I'm thinking of the north pole of a sphere, and the Greenwich meridian passing through it, which cuts out a great circle on the sphere. There are all sorts of smaller circles tangent to that great circle at the north pole, and an extremely small circle has *enormous* curvature!

Teacher. That's true; however, the *normal component* of the curvature vector of each little circle always has the same length. Figure 14.1 shows the idea:

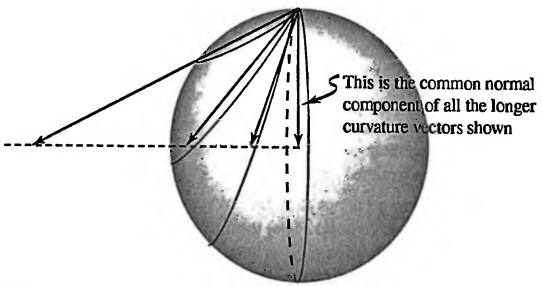


FIGURE 14.1. Planes through a point on a sphere intersect the sphere in circles, and the closer any plane comes to being tangent to the sphere, the smaller the circle. Of course, that means that the curvature vector gets longer. However, its *normal component* remains constant. This turns out to be true of any smooth surface, and this fact is the essential claim in "Meusnier's Theorem," which he discovered in 1776.

Student. When do we apply all this to the complex unit circle?

Teacher. Right now. Let S be the complex circle $x^2 + y^2 = 1$, and for concreteness, let P be the point $(1, 0) \in \mathbb{C}^2$—that is, $(1, 0, 0, 0) \in \mathbb{R}^4$. The tangent plane $\mathbf{T}$ at P is the complex line $x = 1$ in $\mathbb{C}^2 = \mathbb{R}^4$. Figure 14.2 shows it sitting in a 3-space of $\mathbb{R}^4$:

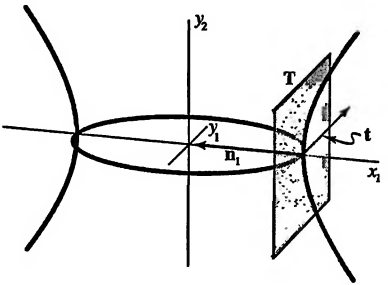


FIGURE 14.2. The shaded plane $\mathbf{T}$ is the complex line containing the real tangent line to the real circle $x^2 + y^2 = 1$ at $(1, 0)$.

The plane normal to S at $(1, 0, 0, 0)$ has equation $y = y_1 + iy_2 = 0$, so this normal plane is just the (x_1, x_2)-plane. Within the confines of a 3-D slice, we can show only one of its two dimensions. In Figure 14.2, this is the x_1-axis. Let $\mathbf{n}_1$ be the left-pointing unit normal on the x_1-axis.

With this setup, what does the "sweeping through the unit tangent vectors" scenario look like? Let's start with the unit tangent vector $\mathbf{t} = (0, 0, 1, 0)$, which lies on the y_1-axis. The plane containing this and $\mathbf{n}_1$ is the (x_1, y_1)-plane. We immediately encounter a pleasant exception, for this plane intersects S in a curve (the real circle) instead of just the point P. The normal component of this circle's curvature vector is in this case just $\mathbf{n}_1$.

Now within the tangent plane $\mathbf{T}$, rotate $\mathbf{t}$ a bit to $\mathbf{t}^*$. Here's the basic idea: the plane containing $\mathbf{n}_1$ and $\mathbf{t}^*$ intersects S in only a point; it turns out that the normal component of the curvature vector to any curve on S with tangent $\mathbf{t}^*$ is just a little shorter than $\mathbf{n}_1$. When $\mathbf{t}$ rotates through an angle θ in $\mathbf{T}$, the normal component of the curvature vector is $-\cos\theta$. Its extremes are, of course, -1 and $+1$ and their product is -1. Notice that the $+1$ is attained when $\theta = \pi$, in which case the tangent vector is $(0, 0, 0, 1)$. That turns out to be the other instance where we actually have an intersection curve—the hyperbola opening up to the right in the figure. Relative to $\mathbf{n}_1$, the curvature of this real hyperbola at P is $+1$. Amazingly, we actually see precisely the two cases where the curvature is extreme.

Student. When I look at Figure 13.1 [p. 235] and see the circle bending one way and the attached hyperbola bending the other way, it doesn't stretch my imagination too much to guess that we've got a saddle-surface of some sort. And when we were talking earlier about the Gauss map, we saw that on a saddle surface, the Gauss map reverses orientation and therefore the curvature is regarded as negative. So at least I can accept the negative sign. But...

Teacher. You're still wondering about -2 instead of -1 for the Gausslan curvature at P?

Student. Yes.

Teacher. Everything we have so far carried out is relative only to that one orthonormal basis vector $\mathbf{n}_1$. But the normal plane has *two* orthogonal vectors! Let the other "paint brush" be $\mathbf{n}_2$, pointing along the negative x_2-axis. When $\mathbf{t}$ rotates through θ in $\mathbf{T}$, the $\mathbf{n}_2$-normal component of the curvature vector is now $\sin\theta$. We once again arrive at extreme sectional curvatures of ± 1, and we again get a product -1. The two paintbrushes paint out a total area whose limiting ratio is $-1-1=-2$.

Philosopher. Do we still see a circle and hyperbola in a 3-D slice, as in Figure 14.2?

Teacher. No. The circle equation's real and imaginary parts are

$$x_1^2 - x_2^2 + y_1^2 - y_2^2 = 1 \quad \text{and} \quad x_1x_2 + y_1y_2 = 0.$$

The 3-space in Figure 14.2 has equation $x_2 = 0$ because it's generated by $\mathbf{T}$ and $\mathbf{n}_1$, and the circle-plus-hyperbola is the intersection of the hyperboloid $x_1^2 + y_1^2 - y_2^2 = 1$ with the two planes $y_1y_2 = 0$. Its translate, now generated by T and $\mathbf{n}_2$, is defined by $x_1 = 1$. We therefore have the intersection of

$$-x_2^2 + y_1^2 - y_2^2 = 0 \quad \text{and} \quad x_2 + y_1y_2 = 0,$$

which are respectively a circular cone and hyperbolic paraboloid. This intersection again consists of two space curves "at right angles" in 3-space, but not a circle and hyperbola. One can check that these curves have parametrizations

$$y_1 = t; \quad y_2 = \pm\frac{t^2}{\sqrt{1+t^2}}; \quad x_2 = \mp\frac{t}{\sqrt{1+t^2}}.$$

Applying a standard formula for space curvature (see Struik [13, p 11]) gives curvatures of $+1$ and -1 at $(1, 0, 0, 0)$, and therefore product -1.

Ellipse Curvature and Some Pictures

Philosopher. Can you generalize what we just did from the complex unit circle to complex conics

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1?$$

Teacher. Yes—there are only minor differences. At the point $(a, 0)$, for example, as $\mathbf{t}$ rotates by θ within the plane $\mathbf{T}$, the $\mathbf{n}_1$- and $\mathbf{n}_2$-components of the curvature vectors are no longer $-\cos\theta$ and $\sin\theta$, but rather

$$\frac{-a}{b^2}\cos\theta \quad \text{and} \quad \frac{a}{b^2}\sin\theta.$$

This leads to extrema

$$\pm\frac{a}{b^2},$$

with product $-\frac{a^2}{b^4}$; therefore, the Gaussian curvature at $(a, 0)$ is

$$-\frac{2a^2}{b^4}.$$

Philosopher. By symmetry, the Gaussian curvature at $(0, b)$ is

$$-\frac{2b^2}{a^4},$$

right?

Teacher. Yes. These Gaussian curvatures at the vertices $(a, 0)$ and $(0, b)$ are extremes; the ratio of the first to the second is

$$\frac{a^2}{b^4} \div \frac{b^2}{a^4} = \left(\frac{a}{b}\right)^6.$$

Philosopher. The sixth power! This means that only a small deviation from being a circle could translate to pretty large relative variations in Gaussian curvature.

Teacher. That's right. For example if $a = 1.05$ and $b = 1$, then

$$\left(\frac{a}{b}\right)^6 = 1.05^6 \approx 1.34.$$

Expressed another way, a 5% increase in a results in nearly 7 times the percentage variation in the Gaussian curvature!

Student. I'd like to see a picture of this.

Teacher. Sure. Let's begin with a picture of the Gaussian curvature when $a = 1$, $b = 1$; its equation is

$$\kappa_{\text{unit circle}} = \frac{-2}{\cosh^3 2v},$$

appearing in (14.5) [p. 270], and Maple plots this picture:

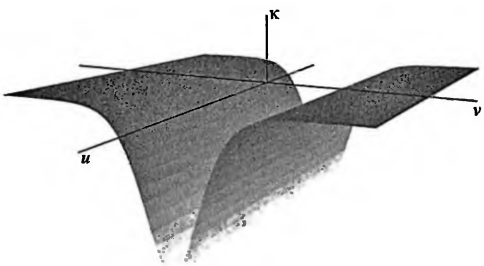


FIGURE 14.3. This is a graph of the Gaussian curvature of the complex unit circle $x^2 + y^2 = 1$. The u-axis runs parallel to the bottom of the trough. That bottom corresponds to $\kappa = -2$, which is the (constant) Gaussian curvature of the complex circle at points of the real circle.

Now an explicit formula for the Gaussian curvature when, say, $a = 1.05$, $b = 1$, would be much more complicated than (14.5). However, Maple is happy to accept fancy formulas like the ones in (13.32) and (13.33), once $\mathbf{r}$ is given. Maple will chug away, creating this nice picture:

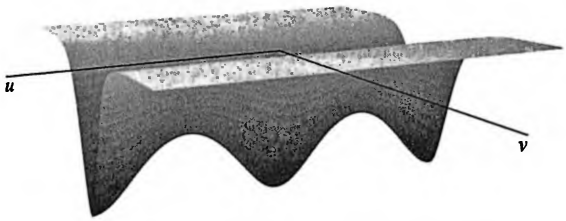


FIGURE 14.4. This is a graph of the Gaussian curvature of the ellipse

$$\frac{x^2}{1.05^2} + \frac{y^2}{1^2} = 1.$$

Though the real ellipse looks nearly circular to our eyes, the ratio

$$\frac{a}{b} = 1.05,$$

when raised to the sixth power, is about 1.34. This sixth power is the ratio of maximum to minimum Gaussian curvatures attained on the real ellipse. In the graph, the Gaussian curvatures of the real ellipse appear directly below the u-axis.

Notice how Figure 14.4 shows that the Gaussian curvature indeed varies as one travels along the real ellipse

$$\frac{x^2}{1.05^2} + \frac{y^2}{1^2} = 1.$$

The graph of that curvature lies directly above the u-axis.

Philosopher. That's impressive. Unless you look closely, the real ellipse with $a = 1.05$ and $b = 1$ appears to be a circle.

Philosopher. Even nudging a a little more, say to 1.5, would further accentuate what we see in Figure 14.4.

Teacher. Yes. So much so, you see deep troughs:

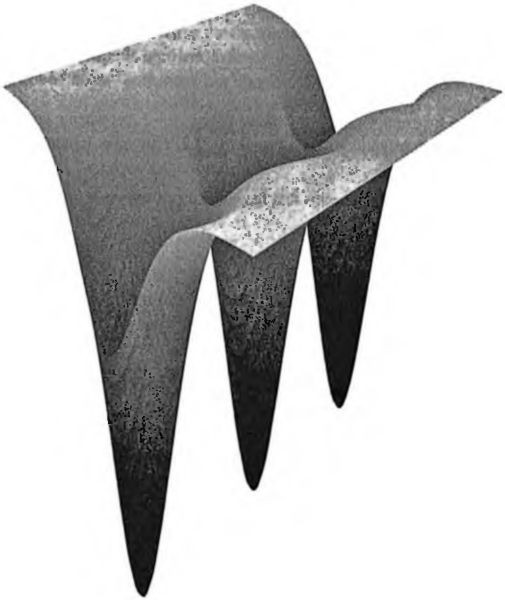


FIGURE 14.5. Here's a plot of the Gaussian curvature for the ellipse

$$\frac{x^2}{1.5^2} + \frac{y^2}{1^2} = 1.$$

The curvature varies so much that we see a series of deep troughs below the u-axis. This behavior quickly calms down as v moves away from zero.

For a 10-by-1 ellipse, the Gaussian curvature along its cigar-shaped real part varies by exactly a million-to-one ratio.

Philosopher. A number of these pictures suggest that for v near zero, the Gaussian curvature can be a show-stopper, but as v grows in size, the curvature subsides to nearly zero.

Interactive Applet

To rotate Figure 14.5, see Chapter Applet 19 on the CD-ROM.

Teacher. I agree. Take the Gaussian curvature

$$\kappa = \frac{-2}{\cosh^3 2v}$$

of the unit circle, for example. When $v = 0$, we have $\kappa = -2$, but when v grows to 3, that curvature decreases in magnitude to less than 0.1% of 2. Thereafter, it quickly approaches its limiting value of $\kappa = 0$.

Student. From the pictures, it seems that the curvature, even when its absolute value is small, is always negative.

Teacher. That's true. Intuitively, a nondegenerate conic is everywhere "saddle-shaped." Figure 13.1, for example, strongly suggests that the unit circle has a saddle-shape around the point (1, 0), since the real circle bends one way and the attached hyperbola bends the opposite way. Although in $\mathbb{R}^4$ it's not trivial to say just what "bending in opposite directions" means, a negative κ lends some algebraic precision to this idea.

As another bit of intuition, complex conics—in fact, complex algebraic curves, in general—are minimal.

Student. What is a "complex algebraic curve"?

Teacher. A *plane* complex algebraic curve is the set of points in $\mathbb{C}^2$ satisfying a polynomial equation $p(x, y) = 0$. One can extend this to $\mathbb{C}^n$ by considering several polynomials whose set of common solutions has complex dimension 1; and by going to its topological closure, we can take the curves to lie in $\mathbb{P}^n(\mathbb{C})$. For example, we stated Bézout's theorem for algebraic curves in $\mathbb{P}^2(\mathbb{C})$.

Student. And what do you mean by "minimal"?

Teacher. A square patch in our real (u, v)-parameter space maps onto a patch of the conic whose boundary B is generally a slightly-wracked square—a square that's been somewhat twisted out of shape. Of all the surfaces in $\mathbb{R}^4$ having B as boundary, the one having least area is called *minimal*—a "soap film surface." The minimal surface with boundary B coincides with the little patch of our complex conic, which has negative curvature at each point.

> You can get a feeling for the idea of "minimal surface" in $\mathbb{R}^3$ by dipping a somewhat bent-out-of-shape wire loop into a soap solution. The resulting soap film has negative curvature at each point—geometrically, it's "saddle shaped" at each point.

Student. I don't see why the curvature of an actual soap film couldn't be positive...

Teacher. Well, the soap film is under tension—"surface tension"—somewhat like the stretched fabric of a trampoline. Now suppose you're jumping on a trampoline. When you land on it you create a bowl-shaped depression, and that has positive curvature. The elastic fibers making up the fabric are eager to spring back to their original shorter lengths, and

you feel that eagerness by being propelled upward. On the other hand, if the trampoline's frame had been twisted out of shape during shipping, you'd see a slightly saddle-shaped surface. Any bowl-shape is transient, because that shape leaves a net unbalanced force in the trampoline fibers—somewhat like an arc-shaped violin string wanting to become straight. But perhaps you've seen string models of negative-curvature surfaces, such as a hyperboloid of one sheet, or a "horse-saddle"; the ends of the straight strings are attached to a frame, and all those straight "violin strings" are happy to remain straight.

Are Hyperbolas Any Different?

Student. We've never said anything about the curvature of complex hyperbolas.

Philosopher. We've been successfully using the point of view that a hyperbola and ellipse are essentially the same, so I bet we've basically done all the work.

Teacher. I agree. As a surface $S \subset \mathbb{R}^4$, a complex ellipse consists of the common solutions to the real and imaginary parts of the complex equation

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1,$$

namely

$$\frac{x_1^2}{a^2} - \frac{x_2^2}{a^2} + \frac{y_1^2}{b^2} - \frac{y_2^2}{b^2} = 1; \quad \frac{x_1 x_2}{a^2} + \frac{y_1 y_2}{b^2} = 0. \tag{14.6}$$

The corresponding equations for the complex hyperbola

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$$

are

$$\frac{x_1^2}{a^2} - \frac{x_2^2}{a^2} - \frac{y_1^2}{b^2} + \frac{y_2^2}{b^2} = 1; \quad \frac{x_1 x_2}{a^2} - \frac{y_1 y_2}{b^2} = 0. \tag{14.7}$$

Now visualize the surface S defined by (14.6) sitting in $\mathbb{R}^4$, with $\mathbb{R}^4$ equipped with labels x_1, x_2, y_1, y_2 on the four coordinate axes. Go to the y_1-label and change its name to y_2. Go to the original y_2-label, move it to the negative side of its axis, and change its name to y_1.

Philosopher. You've made the substitutions

$$y_1 \to y_2,$$

$$y_2 \to -y_1.$$

Teacher. Yes. Notice that this does absolutely nothing to the surface S itself. What we've done is no different from changing street signs, which doesn't do anything to the houses on the street! Of course, the curvature at a given point of S hasn't changed, either. So if $f(x_1, x_2, y_1, y_2)$ is a formula for the curvature of the ellipse in the original coordinate system, then by changing coordinate-names—the street signs—we get $f(x_1, x_2, y_2, -y_1)$ as the formula for the curvature of the corresponding hyperbola.

Student. That seems almost too easy! How would you apply this to find the curvature of the hyperbola

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$$

at the vertex $(a, 0)$?

Teacher. In $\mathbb{R}^4$, this vertex has coordinates $(a, 0, 0, 0)$, and the ellipse has Gaussian curvature

$$f(a, 0, 0, 0) = \frac{-2a^2}{b^4}$$

there. The Gaussian curvature of this same point on the hyperbola is therefore

$$f(a, 0, 0, -0) = f(a, 0, 0, 0) = \frac{-2a^2}{b^4},$$

unchanged.

Student. What about other points on the hyperbola?

Teacher. Our idea of changing street signs means that the ellipse parametrization

$$\mathbf{r}(u, v) = (a \cos u \cosh v, -a \sin u \sinh v, b \sin u \cosh v, b \cos u \sinh v) \tag{14.8}$$

becomes

$$\mathbf{r}(u, v) = (a \cos u \cosh v, -a \sin u \sinh v, b \cos u \sinh v, -b \sin u \cosh v), \tag{14.9}$$

a parametrization of the hyperbola. When we substitute this into our dot-product curvature formula (13.32) [p. 265], that formula doesn't change at all.

Student. That is amazing! Sort of like magic.

Teacher. It's actually quite simple. For example, the dot product $(1, 2, 3, 4) \cdot (1, 2, 3, 4)$ is the same as $(1, 2, 4, 3) \cdot (1, 2, 4, 3)$—any consistent switching of coordinates doesn't affect the dot product. And this last product is the same as $(1, 2, 4, -3) \cdot (1, 2, 4, -3)$; introducing the minus sign doesn't do anything either, since $(-1)^2 = 1$. In addition, all minus signs survive differentiations. This means that all entries in (13.32) are unchanged.

Philosopher. That means those curvature pictures, Figures 14.3–14.5, are unchanged!

Teacher. That's right.

Student. How could that be?! When I look at an ellipse, I see curvature changing between a max and a min, but with a hyperbola, I see the curvature getting less and less the farther out you go.

Teacher. Although the surface S in $\mathbb{R}^4$ is unaffected when we change coordinates, the real part—just a real 2-D view of S—does change. So the ellipse and hyperbola are simply different, very limited views, of the same "elephant." With the ellipse, we're observing the u-curve $v = 0$. With the hyperbola, we're observing the v-curve $u = 0$. This is reflected in the corresponding parametrizations: when $v = 0$, the ellipse in (14.8) has the real parametrization $(a \cos u, 0, b \sin u, 0)$; when $u = 0$, the hyperbola in (14.9) has the real parametrization $(a \cosh u, 0, b \sinh u, 0)$.

Student. We obtained a curvature formula

$$\kappa = \frac{-2}{\cosh^3 2v}$$

for the complex unit circle $x^2 + y^2 = 1$. Are you telling me that this same formula works for the corresponding complex hyperbola $x^2 - y^2 = 1$?

Teacher. Yes. On the real circle, v is always 0, while on the real hyperbola it is u which remains 0, while v varies. So we actually see the Gaussian curvature varying on the real hyperbola.

Student. So the curvature of the real hyperbola is

$$\frac{-2}{\cosh^3 2v}?$$

Teacher. No—not the *real* hyperbola. Remember when we considered the unit circle? At its real points, the complex circle S has constant Gaussian curvature -2. In a sense we pick up half of this in the curvature of the corresponding real curve. Does a similar thing happen with the hyperbola? We can easily find out, because we've worked out a general formula, (13.1), for the curvature of a parametrized plane curve:

$$\kappa = \frac{\dot{x}\ddot{y} - \ddot{x}\dot{y}}{((\dot{x})^2 + (\dot{y})^2)^{3/2}}.$$

Using the real hyperbola's parametrization $x = \cosh v$, $y = \sinh v$ and substituting into our formula gives

$$\frac{\dot{x}\ddot{y} - \ddot{x}\dot{y}}{((\dot{x})^2 + (\dot{y})^2)^{3/2}} = \frac{\sinh^2 v - \cosh^2 v}{(\sinh^2 v + \cosh^2 v)^{3/2}} = \frac{-1}{\cosh^3 2v},$$

which is again just half of the Gaussian curvature of the surrounding surface!

Chapter Summary

This chapter begins by applying the last chapter's surface-curvature formula in (13.32) and (13.33) on p. 265 to obtain the curvature of a complex circle having the parametrization in (14.1) on p. 269. The result is

$$\kappa = \frac{-2}{\cosh^3 2v}.$$

The real circle $x^2 + y^2 = 1$ corresponds to $v = 0$, and that means the surface curvature on the real circle is -2, not 1; we provide some geometric insight for the difference. The formula for a noncircular ellipse is more complicated, but Maple can easily generate plots that are both beautiful and informative; these appear in Figures 14.3–14.5. To see a rotatable picture of Figure 14.5, see Chapter Applet 19 on the CD-ROM. The chapter concludes by showing that the substitution $y_1 \to y_2$, $y_2 \to -y_1$ changes the curvature formula for the complex ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

into that of the hyperbola

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1.$$

Exercises

1. We observed in (14.4) on p. 270 that when $a = b = 1$, (13.32) on p. 265 simplifies quite a bit, leading to (14.5) for the Gaussian curvature of a complex circle. From our "street sign" argument, it would seem that taking $a = 1, b = i$ in (13.32) should lead to a similar simplification. Carry out the computations to see if this is in fact so.
2. Prove Meusnier's theorem, illustrated in Figure 14.1 on p. 272, for an ordinary real sphere of radius R.
3. If k is a nonzero real number, how do you modify the curvature-surface graph of

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

to get the curvature-surface of

$$\frac{x^2}{(ka)^2} + \frac{y^2}{(kb)^2} = 1?$$

How, therefore, does the curvature-surface in Figure 14.3 on p. 275 change as you vary the circle's radius?

4. Find the Gauss curvature of the complex parabola $y = ax^2$. For various real choices of a, use Maple to generate some curvature-surfaces similar to the ones in Figures 14.3–14.5.
5. We know that a circle of pure imaginary radius (negative radius-squared) defines a card-carrying real surface in $\mathbb{R}^4$ What is its Gaussian curvature? Plot an analogue of Figure 14.3.
6. Although we've used the power of the *Theorema egregium* to define surface curvature in $\mathbb{C}^2 = \mathbb{R}^4$, one can nevertheless do a little reality-checking to see if such a definition makes geometric sense. For example, rotating an ordinary conic in $\mathbb{R}^2$ through some real angle ϕ does nothing to the conic's curvature, nor should it do anything to the curvature of a complex conic. The equations in (A1.5) of Appendix 1 parametrize the ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

after undergoing rotation by ϕ. Using Maple or Mathematica, show that the expression in (13.32), divided by that in (13.33) on p. 265, is the same for this ellipse before and after rotation.

15

Photons and Conics

in which an old question about reflecting light gets an answer

Short Trips, Long Trips

Philosopher. I'm still puzzled about how light reflects off conics. If a burst of photons leaves one focus of a mirrored ellipse and reflects off its walls, the photons all simultaneously meet at the other focus. But with a mirrored hyperbola, every photon heads the wrong way, starting an endless journey into outer space! To make the standard picture work, the photon has to be turned around and forced to move backward so it heads toward the other focus.

> We physicists are always checking to see if there is something the matter with the theory. That's a game, because if there *is* something the matter, it's interesting!
>
> Richard Feynman (QED, p. 8)

Teacher. That definitely doesn't sound like an ideal state of affairs. I will say this, however: after reflecting, a photon does travel on one and the same line, whether you reverse its direction or not. If it goes backwards, it completes its journey in a finite amount of time. If it moves forward, it doesn't. In fact, it can't even make it to the line at infinity in any finite amount of time, and it certainly must cross that to get to the other focus. And the standard device of subtracting distances for the hyperbola? I suspect that is related to the photon's moving backwards.

Philosopher. So for hyperbolas, if a photon travels the wrong way, it's a finite trip, while if it goes the right way, it's an infinite one.

Student. Too bad photons don't travel in circles!

Teacher. Come again?

Student. To travel on a circle from one point to another, you can go the short way or the long way, but either way the distance is *finite*.

Philosopher. So you could fly east on a great circle from San Francisco to New York, getting there "the short way," or you could head west from San Francisco, traveling on that same great circle, and finally make it to New York.

Student. Yes, that's my point.

Philosopher. A tempting idea. However, photons don't travel along circles.

Teacher. But they *do* travel on geodesics, so if a photon's universe were a sphere, then it would travel along a great circle. Come to think of it, I once read that photons can theoretically get trapped just inside the horizon of a black hole, traveling around in circles!

Student. You mean my idea might actually work?

Teacher. I think it's worth exploring. The sphere concept already occupies a central position in our study, the real projective plane being a sphere with opposite points identified. Alternatively, that's an ordinary hemisphere with opposite boundary points identified. What if we define our conic on a hemisphere in the same spirit as before? That is, the curve is the intersection of a cone and a hemisphere with its opposite boundary points identified. As usual, the cone's vertex is at the center of the sphere. Under central projection, lines tangent to the hemisphere map to "great semicircles."

Student. But I was talking about circles, not semicircles.

Teacher. But opposite points of the hemisphere boundary are identified, so any semicircle closes up to give a loop.

Philosopher. A loop, yes. But to make the idea work, measuring distances on that closed semicircle has to be the same as measuring distances on an ordinary circle. Does that actually happen?

Teacher. We've got to *make* it happen! When we earlier considered the disk model of $\mathbb{P}^2(\mathbb{R})$, we transferred the ordinary plane's metric to the disk. And we can define a metric on a hemispherical model using central projection, defining the distance between any two points on the upper hemisphere as the distance between their preimages in the plane tangent at the north pole. The metric on the hemisphere therefore becomes ever more concentrated near the boundary.

But now I'm going to propose a different metric on the hemisphere—*the usual one inherited from sitting in* $\mathbb{R}^3$!

Philosopher. So if a plane flies east from San Francisco to New York along a geodesic, then the distance is the length of the flight path.

Teacher. Yes, and more generally, take a tape measure and pull it taut between two points on the hemisphere so it lies on a great circle—a geodesic. The length of the arc is then whatever the tape measure reads. Of course, there are two ways to go around a circle. Now on the hemisphere, how do you measure the distance from San Francisco to New York if you fly *west*? Well, continue to stretch your tape measure taut along the geodesic until you reach the boundary—the equator. Glue the tape to the equator and then cut the tape measure at that boundary point. Bring the freshly-cut, free tape end over to the opposite point of the equator and continue measuring. What we read off the tape measure in New York is what we can call the "long-trip distance." Simply for convenience, we'll call any trip that crosses the equator. "long:" and any other trip. "short."

Philosopher. I see that your cutting idea makes the distance between opposite boundary points zero! But do these intuitive ideas actually work? For example, what about the sum of distances to the foci? For ordinary ellipses, we know that sum is a constant, but is that true on our hemisphere? And by using the functions $L_1(x)$ and $L_2(x)$ in (5.5) and (5.6) [p. 77], the signed distances continue to add to a constant even for a plane hyperbola, illustrated in Figure 7.6 [p. 114]. Does this work on a hemisphere? And if a photon goes the other way around on the hemisphere, does sum-of-distances still hold?

Teacher. We certainly need to answer such questions. On that last point, just talking intuitively, suppose that for an ellipse on the hemisphere we have a "short trip" relation "$L_1(x) + L_2(x) = K$," where K is a constant. If the circumference of a great semicircle is C, then the alternative routes have lengths $C - L_1(x)$ and $C - L_2(x)$, and these add to

$$\big(C - L_1(x)\big) + \big(C - L_2(x)\big) = 2C - \big(L_1(x) + L_2(x)\big) = 2C - K.$$

Since all great semicircles have the same circumference, $2C - K$ is a constant, too! One could say that "a short-trip sum-of-lengths implies a long-trip sum-of-lengths."

Philosopher. And obviously vice-versa. This approach looks promising. Let's continue exploring it.

The Pythagorean Theorem on a Sphere

Teacher. In the plane, many results depend on computing the distance between two points—for example, getting the distances from the foci to an arbitrary point on a conic, or from the conic's center to a focus or directrix. In these the Pythagorean theorem is a basic tool, and in "doing conics on a hemisphere," one should have a spherical analogue of this theorem. Of course on a sphere, "line segments," such as the sides of a spherical triangle, are actually arcs of great circles.

> Radians give the simplest form to various formulas involving the circle. For example, when θ is measured in radians, then on a circle of radius r, an arc subtending a central angle θ has length $r\theta$. But if θ is in degrees, one has to include the unpleasant factor $\frac{\pi}{180}$, making the arc length $\frac{\pi}{180}r\theta$. A similar thing holds for derivatives involving angles. For example, the derivative of $\sin\theta$ is $\cos\theta$ only when θ is in radians—in degrees, one has to include that same unpleasant factor. The derivative then reads $\frac{\pi}{180}\cos\theta$. The familiar power series expansions of trigonometric functions once again assume that the variable is measured in radians. Only then can we write an equation like
>
> $$\sin\left(\frac{\pi}{2}\right) = 1 = \frac{\pi}{2} - \frac{1}{3!}\left(\frac{\pi}{2}\right)^3 + \frac{1}{5!}\left(\frac{\pi}{2}\right)^5 - \frac{1}{7!}\left(\frac{\pi}{2}\right)^7 + \cdots.$$

The angle between two arcs intersecting at a point P is the angle between their tangent lines there. Thus, for example, the terms "right triangle" and "hypotenuse" make sense on a sphere. Here is a spherical version of the Pythagorean Theorem:

Spherical Pythagorean Theorem. Let A, B and C be the legs and hypotenuse of a right triangle on a sphere of radius R. Then

$$\cos\frac{C}{R} = \cos\frac{A}{R} \cdot \cos\frac{B}{R}. \tag{15.1}$$

Proof Let's write

$$\alpha = \frac{A}{R}, \quad \beta = \frac{B}{R} \quad \text{and} \quad \gamma = \frac{C}{R}.$$

Then establishing (15.1) is the same as proving that $\cos\gamma = \cos\alpha \cdot \cos\beta$. Now α, β and γ are the corresponding sides of a right triangle on a *unit* sphere. From the relation $s = r \cdot 1$, they are also the central radian angles subtended by these sides. This is illustrated in Figure 15.1.

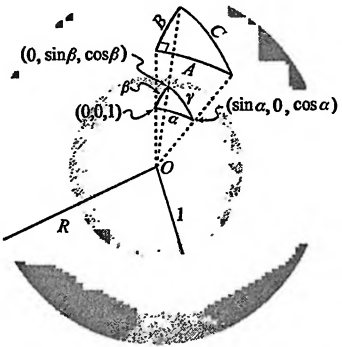


FIGURE 15.1. Triangle ABC subtends central angles α, β and γ, and when measured in radians, they are numerically equal to the corresponding sides of the triangle's central projection on the unit sphere.

The unit sphere—the smaller one in Figure 15.1—has equation $x^2+y^2+z^2 = 1$. We've chosen x- and y-axes in $\mathbb{R}^3$ so that the legs α and β are in the planes $y = 0$ and $x = 0$, respectively. The legs therefore intersect at (0, 0, 1). As shown in the figure, the space-coordinates of the other two vertices are $v = (\sin\alpha, 0, \cos\alpha)$ and $w = (0, \sin\beta, \cos\beta)$. Since the angle between these two unit vectors is γ, their dot product $v \cdot w = \|v\|\|w\| \cos\gamma$ becomes

$$0 + 0 + \cos\alpha \cdot \cos\beta = 1 \cdot 1 \cdot \cos\gamma,$$

which is what we wanted to prove.

Student. That's a neat theorem, but I would never have guessed it had anything to do with the usual Pythagorean theorem!

Philosopher. You can regard a plane as a sphere of infinite radius, so it seems to me that as $R \to \infty$, the spherical version ought to tend toward the usual plane version.

Teacher. And it does! The secret to seeing this is to expand (15.1)'s cosines in Maclauren series:

$$1 - \frac{C^2}{2!R^2} + \frac{C^4}{4!R^4} - \cdots = \left(1 - \frac{A^2}{2!R^2} + \frac{A^4}{4!R^4} - \cdots\right) \cdot \left(1 - \frac{B^2}{2!R^2} + \frac{B^4}{4!R^4} - \cdots\right).$$

In this equation, multiply out the right side, subtract 1 from each side, then multiply through by $-2R^2$. The result can then be written as

$$C^2 = A^2 + B^2 + \textit{Remainder},$$

where each term of the bounded *Remainder* has a positive power of R in the denominator. One can then check that *Remainder* $\to 0$ as $R \to 0$, leaving the usual Pythagorean theorem.

Philosopher. It amazes me how things that superficially seem as far apart as the plane and sphere versions of the Pythagorean theorem are actually closely related.

Student. It's your Hawaiian Islands philosophy again!

Stakes and Strings on a Sphere

Philosopher. Now that we have a mechanism for computing distances on a sphere, we should see if conics on a hemisphere act the way we want.

Teacher. Right. Let's assume our hemisphere is defined by $x^2 + y^2 + z^2 = R^2$, where $z \geq 0$, and our elliptical cone, by

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = \frac{z^2}{R^2}.$$

The plane $z = R$ is tangent to the hemisphere at its north pole $(0, 0, R)$, and the cone intersects this plane in the ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = \frac{R^2}{R^2} = 1.$$

As a first application of our spherical Pythagorean theorem, let's obtain the spherical analogue of the traditional focal length $f = \sqrt{a^2 - b^2}$ for this ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1.$$

For definiteness, assume $a > b$.

Let's denote by A and B the central projections of the semi-axes a and b into the R-hemisphere. The idea is shown in the edge-on view of Figure 15.2:

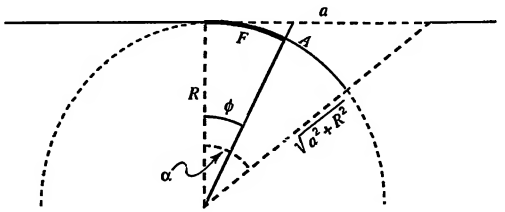


FIGURE 15.2. The semi-major axis a of an ellipse in the plane $z = R$ centrally projects to the solidly-drawn arc A subtending angle α. This arc will be moved so as to become the hypotenuse of a spherical right triangle. The ellipse semi-minor axis b is perpendicular to this page, so we don't see it. B, also not shown, is its analogous projection. The "focal arc" F, drawn more heavily, subtends central angle ϕ, where

$$\cos\phi = \frac{\cos\alpha}{\cos\beta}.$$

One can slide A around so it joins endpoints of B and F to form a spherical right triangle.

a is the length of the horizontal dashed line segment and spherical triangle side A is the solid arc on the R-hemisphere subtending the angle α. In analogy to the ordinary plane case, regard the "spherical focal arc" F as a leg of a spherical right triangle $\{A, B, F\}$, where A is the hypotenuse. Let

$$\alpha = \frac{A}{R}, \quad \beta = \frac{B}{R}, \quad \phi = \frac{F}{R}.$$

Now in a plane, the x-focal length is $\sqrt{a^2 - b^2}$—that is, it's computed using the plane Pythagorean theorem. Figure 6.2 [p. 94] shows the idea. On a sphere, the analogous length is computed using the spherical Pythagorean theorem. That is,

$$\cos\phi = \frac{\cos\alpha}{\cos\beta}. \tag{15.2}$$

Student. If conics on a hemisphere act as we're hoping they do, then the "sum-of-lengths" property should also hold, right?

Teacher. Yes. The ordinary sum-of-lengths constant is the long diameter of the plane ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1.$$

The hemispherical analogue is the length of the corresponding great-circle arc. We see from Figure 15.2 that this is

$$2R\text{Arctan}\left(\frac{a}{R}\right). \tag{15.3}$$

Student. Does that mean you're going to pick an arbitrary point on the ellipse and see if the great-circle arc lengths add to $2R\text{Arctan}\left(\frac{a}{R}\right)$? I'd like to see if that works for a specific example!

> A photon, just like any fundamental particle, has both a particle and a wave nature. But whether an ideal ellipse reflects billiard balls or waves, if they emanate from one focus, they converge to the other. And if unfettered, they both continue on, alternately converging first at one focus, then at the other. But does a wave actually remain a wave? (Try running Chapter Applet 21 on the CD-ROM.)

Teacher. Good idea. Let's choose a not-too-special example, like

$$\frac{x^2}{3^2}+\frac{y^2}{2^2}=1$$

in the tangent plane $z = 5$ of the hemisphere with radius $R = 5$. In this case, the long diameter is $2R\text{Arctan}\left(\frac{a}{R}\right)$—that is,

$$10\text{Arctan}\left(\frac{3}{5}\right) = \text{purported spherical ellipse constant.}$$

A somewhat "arbitrary" point in this plane ellipse is

$$\left(1, \frac{4\sqrt{2}}{3}\right).$$

This point has space coordinates

$$\left(1, \frac{4\sqrt{2}}{3}, 5\right),$$

and R times the normalization of this vector is its central projection to the R-hemisphere. This central projection turns out to be

$$\mathbf{v} = \left(\frac{15}{\sqrt{266}}, \frac{20}{\sqrt{133}}, \frac{75}{\sqrt{266}}\right). \tag{15.4}$$

Student. I'll bet you're going to do the same thing for the foci...

Teacher. Yes, let's also get their space coordinates on the hemisphere. The focal length is $R\phi$, where

$$\cos\phi = \frac{\cos\alpha}{\cos\beta}.$$

From Figure 15.2, we see that

$$\alpha = \text{Arccos}\frac{R}{\sqrt{a^2+R^2}};$$

similarly,

$$\beta = \text{Arccos}\frac{R}{\sqrt{b^2+R^2}}.$$

Therefore

$$\cos\phi = \frac{R}{\sqrt{a^2+R^2}} \Big/ \frac{R}{\sqrt{b^2+R^2}} = \sqrt{\frac{b^2+R^2}{a^2+R^2}}.$$

Using $\sin\phi = \sqrt{1-\cos^2\phi}$ leads to

$$\sin\phi = \sqrt{\frac{a^2-b^2}{a^2+R^2}},$$

so the right and left foci on the R-hemisphere are

$$\left(\pm R\sqrt{\frac{a^2-b^2}{a^2+R^2}},\ 0,\ R\sqrt{\frac{b^2+R^2}{a^2+R^2}}\right).$$

With our choice of a, b and R, the corresponding vectors are

$$\mathbf{f_L} = \left(-5\sqrt{\frac{5}{34}},\ 0,\ 5\sqrt{\frac{29}{34}}\right);\quad \mathbf{f_R} = \left(5\sqrt{\frac{5}{34}},\ 0,\ 5\sqrt{\frac{29}{34}}\right). \tag{15.5}$$

Philosopher. So we now have three vectors $\mathbf{v}$, $\mathbf{f_L}$ and $\mathbf{f_R}$, each based at the origin, each 5 units long, with their tips at the two foci and an arbitrary point of the hemispherical ellipse.

Teacher. That's right. The left spherical string length, which we denote by L_1, is R times the angle between $\mathbf{v}$ and $\mathbf{f_L}$ and it's easy to see that this is

$$L_1 = R\text{Arccos}\left(\frac{\mathbf{v}\cdot\mathbf{f_L}}{R^2}\right).$$

Similarly, the right spherical string length L_2 is

$$R\text{Arccos}\left(\frac{\mathbf{v}\cdot\mathbf{f_R}}{R^2}\right).$$

Substituting (15.4) and (15.5) into these gives

$$\begin{aligned} L_1 &= 5\text{Arccos}\left(\frac{3\sqrt{5}}{\sqrt{266}\sqrt{34}}(\sqrt{5\cdot 29}-1)\right);\\ L_2 &= 5\text{Arccos}\left(\frac{3\sqrt{5}}{\sqrt{266}\sqrt{34}}(\sqrt{5\cdot 29}+1)\right). \end{aligned} \tag{15.6}$$

If sum-of-lengths holds, then L_1 and L_2 should add up to $10\,\text{Arctan}\left(\frac{3}{5}\right)$.

Student. But look at all those weird numbers! I don't see how $L_1 + L_2$ could *ever* reduce to $10\,\text{Arctan}\left(\frac{3}{5}\right)$.

Teacher. Let's see what Maple can do. Type the following lines into a Maple worksheet:

```
L1:=5*arccos(3*sqrt(5)/sqrt(266*34)*(sqrt(5*29) - 1));
L2:=5*arccos(3*sqrt(5)/sqrt(266*34)*(sqrt(5*29) + 1));
simplify(10*arctan(3/5) - (L1 + L2));
```

The last line produces **0** as the result! From this, we can see that sum-of-lengths does hold in this example.

Student. I can't believe this! I mean, it's just incredible. My idea is working!

Distance Formulas on a Sphere

Philosopher. I must say this is very encouraging, and exciting, too. But we still need a general proof! How should we proceed?

Teacher. Why not try to mimic what we did earlier for plane ellipses? Remember, we introduced distance functions from the foci to points (x, y) on the ellipse. Take

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1,$$

for example. The distance from the left focus to (x, y) is

$$L_1(x) = a + \frac{\sqrt{a^2 - b^2}}{a} \cdot x,$$

and the distance from the right focus to (x, y) is

$$L_2(x) = a - \frac{\sqrt{a^2 - b^2}}{a} \cdot x.$$

It follows at once that $L_1(x) + L_2(x) = 2a$. It would be great if we could capitalize on this idea.

> Suppose that from one focus of an ellipse, you shoot an ideal billiard ball toward any point of the ellipse. If we assume an ideal reflecting surface and no loss of energy, the ball will forever pass alternately through one focus, then the other. The remarkable fact is that in the limit, the ball will simply move back and forth along the major axis! Two things can slow down convergence to this "steady state":
>
> - If the ellipse is nearly circular (e is close to 0);
> - If from one focus we shoot nearly, but not exactly, toward the other focus.
>
> If either or both of these happen, then for a time the behavior can get further away from motion along the major axis. But the long-term behavior is always motion along the major axis. Of course if the ellipse is exactly a circle, the foci are at the center, so our claim becomes trivially true.

Interactive Applet

For an animated, interactive demonstration of the main points in the above box, see Chapter Applet 20 on the CD-ROM.

Philosopher. But on a sphere?! How could things be linear?

Teacher. I'm thinking more in terms of getting the distance function on the tangent plane $z = R$ and then centrally projecting to the sphere. If the spherical ellipse is defined by the cone

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = \frac{z^2}{R^2}$$

then, as we've noted before, the corresponding planar ellipse has equation

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1.$$

Philosopher. Are you saying that on the tangent plane, you'll try mimicking

$$L_i(x) = a \pm \frac{\sqrt{a^2 - b^2}}{a} \cdot x?$$

Teacher. I'm going to try it. Let's start by observing that each of the above L_i consists of two terms: the semi-axis a, which we see in Figure 15.2; and a term linear in x whose coefficient is chosen so that when x has the directrix value

$$\frac{a^2}{\sqrt{a^2 - b^2}},$$

$L_i(x)$ is either 100% or 0% of the full x-diameter $2a$—that is, the constant c for which

$$a \pm c \cdot \frac{a^2}{\sqrt{a^2 - b^2}}$$

is either 0 or $2a$. This amounts to choosing

$$c = \frac{\sqrt{a^2 - b^2}}{a}.$$

I'm going to use the same philosophy on the sphere.

Philosopher. I see that x doesn't actually appear in Figure 15.2, but it is on the x-axis of the tangent plane, shown edge-on there, right?

Teacher. Yes. Of course, x actually determines *two* points on the spherical ellipse, but by symmetry, the string lengths from the foci to each of the two points are the same, so we needn't worry about that.

Philosopher. I guess I never quite realized it, but when $x = 0$, then by symmetry the two string segments have the same length—each is just half of the diameter $2a$.

Teacher. Right. And we've seen that the spherical diameter of the spherical ellipse in the x-direction is $2R\text{Arctan}\frac{a}{R}$, so for the term corresponding to the planar semi-axis a, I'm going to choose $R\text{Arctan}\frac{a}{R}$. And as we just saw, the second term in the planar version is cx, where

$$c = \frac{\sqrt{a^2 - b^2}}{a}.$$

That's the same as

$$\frac{\text{Focal Length}}{a} \cdot x,$$

so I'll try the following as an analogous second term:

$$R\text{Arctan}\left(\frac{\text{Focal Length}}{a} \cdot x\right) = R\text{Arctan}\left(\sqrt{\frac{a^2 - b^2}{b^2 + R^2}} \cdot \frac{x}{a}\right).$$

Philosopher. All in all, then, you're proposing these as string-length formulas on an R-sphere:

$$L_1(x) = R\text{Arctan}\frac{a}{R} + R\text{Arctan}\left(\sqrt{\frac{a^2 - b^2}{b^2 + R^2}} \cdot \frac{x}{a}\right); \tag{15.7}$$

$$L_2(x) = R\text{Arctan}\frac{a}{R} - R\text{Arctan}\left(\sqrt{\frac{a^2 - b^2}{b^2 + R^2}} \cdot \frac{x}{a}\right), \tag{15.8}$$

right?

Teacher. Yes. Of course we have to prove that these actually work! However, the strategy itself seems pretty simple: translate the geometrically-derived string length into algebra, and compare that algebraic expression with our formula. If there's a devil here, it'll be in the details.

Now in the tangent plane $z = R$, the ellipse has equation

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1.$$

Let (x, y, R) be an arbitrary point of this ellipse in that tangent plane, and consider (x, y, R) as a vector based at the origin. Another vector,

$$\left(-\sqrt{a^2 - b^2}, 0, \sqrt{b^2 + R^2}\right),$$

points to the left focus. We can get the cosine of the angle between these two vectors using the dot-product. Then R times that angle is the spherical string length. This should agree with (15.7). Similarly,

$$\left(\sqrt{a^2 - b^2}, 0, \sqrt{b^2 + R^2}\right)$$

points to the other focus, and a similar approach should lead to (15.8).

Philosopher. OK, I understand your strategy, but...

How to turn a wave into a particle

Simultaneously release a bunch of photons at one focus of an ellipse. They'll start off in all directions, looking like an expanding circular wave. But each one, after a number of reflections, will be traveling near the major axis. Furthermore, by the sum-of-lengths property of an ellipse, each photon takes the same time to make a full round-trip, so we end up with all the photons traveling along the major axis as a closely-knit group.

Animation

To see that a circular burst of photons released from an ellipse focus acts more like a point after a few reflections, see Chapter Applet 21 on the CD-ROM.

Teacher. I admit that doing this by hand would be the mathematical equivalent of swimming in peanut butter, but I'm going to let Maple do the lion's share of the computation. Let's work with the distance function L_1 in (15.7); arguments for L_2 are similar.

We begin by translating the string length to algebra. We can without loss of generality choose the y-coordinate of the arbitrary ellipse point (x, y, R) to be

$$y = +\frac{b\sqrt{a^2 - x^2}}{a}.$$

Dot-multiplying the normalized arbitrary-point vector and our left-focus vector gives, according to Maple,

$$\mathbf{C} = \frac{R\sqrt{R^2 + b^2} - x\sqrt{a^2 - b^2}}{\sqrt{a^2 + R^2} \cdot N} \tag{15.9}$$

where

$$N = \sqrt{x^2 + \frac{b^2(a^2 - x^2)}{a^2} + R^2}$$

is the length of the vector (x, y, R). Thus $\mathbf{C}$ is the Cosine of the angle subtended by the string from the left spherical focus to the arbitrary point (x, y, R) on the spherical ellipse.

Now let's turn to our distance formula conjectures. We of course hope that it agrees with the actual spherical string length just obtained directly from the geometry. First divide (15.7) through by R, obtaining

$$\frac{L_1(x)}{R} = \text{Arctan}\frac{a}{R} + \text{Arctan}\left(\sqrt{\frac{a^2-b^2}{b^2+R^2}}\cdot\frac{x}{a}\right), \tag{15.10}$$

and write

$$\theta = \text{Arctan}\frac{a}{R}, \qquad \psi = \text{Arctan}\left(\sqrt{\frac{a^2-b^2}{b^2+R^2}}\cdot\frac{x}{a}\right)$$

Now apply the sum-of-angles formula

$$\tan(\theta+\psi) = \frac{\tan\theta + \tan\psi}{1 - \tan\theta\cdot\tan\psi}$$

to (15.10). This easily simplifies to

$$\mathbf{T} = \frac{a\sqrt{R^2+b^2} + R\sqrt{a^2-b^2}\cdot\left(\frac{x}{a}\right)}{R\sqrt{R^2+b^2} - x\sqrt{a^2-b^2}};$$

this is the Tangent of the angle subtended by the spherical arc that our conjectured formula produces—that is, subtended by the purported length of the arc from the left focus to our arbitrary point on the spherical ellipse. Now cosine and tangent are related by

$$\cos^2\theta = \frac{1}{1+\tan^2\theta},$$

so in our case, we can pose the question this way:

$$\mathbf{C}^2 - \frac{1}{1+\mathbf{T}^2} \stackrel{?}{=} 0.$$

That is, we hope that **C** (coming from the *actual* string length) leads to the same angle as **T** (coming from our conjectured formula).

Philosopher. You're going to let Maple work on this?

Teacher. Yes. After entering into Maple our expressions for **C** and **T**, enter the line

$$\textbf{simplify}(\mathbf{C}^2 - \mathbf{1}/(\mathbf{1}+\mathbf{T}^2), \textbf{symbolic});$$

In less than a second, Maple produces **0** as the result.

Philosopher. More wonderful news! And everything you've done applies to L_2, as well.

Student. What you've done looks very impressive, but I'd feel better seeing some more reality checks. For example, what do the graphs of L_1 and L_2 look like? And isn't it true that as R gets larger and larger, the situation should look more like that in the plane? For example, is it true that as $R \to \infty$, the graphs look more and more like those of our old L_1 and L_2?

Teacher. We should indeed check to see that our formulas behave in such believable ways.

Let's begin by applying our formulas to the example we considered a little while ago:

$$\frac{x^2}{3^2}+\frac{y^2}{2^2}=1$$

in the tangent plane $z = 5$ to the hemisphere of radius $R = 5$. We've seen that for this example, the long diameter is $10\text{Arctan}(\frac{2}{5})$—the string-length constant. From (15.4) the "somewhat arbitrary" point on this plane ellipse is

$$\mathbf{v}=\left(\frac{15}{\sqrt{266}},\frac{20}{\sqrt{133}},\frac{75}{\sqrt{266}}\right).$$

With $a = 3, b = 2, x = 1$ and $R = 5$, our formulas (15.7) and (15.8) give

$$L_1(1)=5\text{Arctan}\left(\frac{3}{5}\right)+5\text{Arctan}\left(\sqrt{\frac{5}{29}}\cdot\frac{1}{3}\right);$$

$$L_2(1)=5\text{Arctan}\left(\frac{3}{5}\right)-5\text{Arctan}\left(\sqrt{\frac{5}{29}}\cdot\frac{1}{3}\right). \qquad (15.11)$$

Maple's "simplify" command immediately shows that our earlier expressions for the left and right string lengths coincide with $L_1(1)$ and $L_2(1)$, as we now see: first, enter into Maple the expressions in (15.6):

```
LL:=5*arccos(3*sqrt(5)/sqrt(266*34)*(sqrt(5*29) - 1));
RL:=5*arccos(3*sqrt(5)/sqrt(266*34)*(sqrt(5*29) + 1));
```

Next type in the new expressions $L_1(1)$and $L_2(1)$:

```
L11:=5*arctan(3/5) + 5*arctan(sqrt(5/29)/3));
L21:=5*arctan(3/5) - 5*arctan(sqrt(5/29)/3));
```

Then both

```
simplify(L11 - LL);
```

and

```
simplify(L21 - RL);
```

produce **0**.

These last two results of course mean that our formulas do yield the same string lengths as those obtained geometrically for our somewhat arbitrary point.

Philosopher. Excellent news!

Teacher. It is also illuminating to let x vary in this example. The resulting plot looks like this:

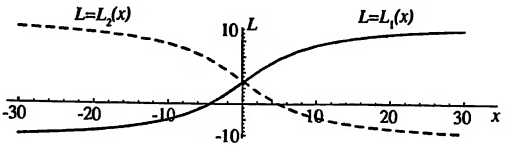


FIGURE 15.3. The solid and dashed curves are the graphs of the left and right string length functions on a sphere, analogous to the left and right string length functions

$$L_1(x) = a + \frac{\sqrt{a^2 - b^2}}{a} \cdot x \quad \text{and} \quad L_2(x) = a - \frac{\sqrt{a^2 - b^2}}{a} \cdot x$$

in the plane.

Student. What happens as $R \to \infty$?

Teacher. Here's a picture showing plots of $L_1(x)$ and $L_2(x)$ for $a = 3$, $b = 2$ and $R = \frac{1}{2}$, 1, 2, 5, 10, 20 and 100:

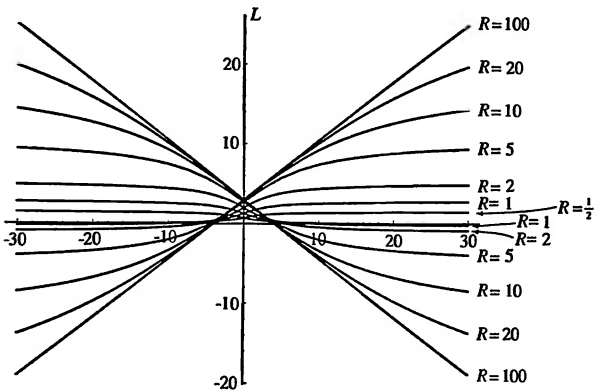


FIGURE 15.4. The graphs of left and right string lengths measured on spheres of increasing radii. Note that as R grows, the graphs approach the corresponding left and right straight-line graphs for a plane ellipse.

One can clearly see the graphs approaching those of the ordinary planar case as R increases.

Philosopher. Can we justify what Figure 15.4 suggests by working with the formulas in (15.7) and (15.8)?

Teacher. Yes, we can. Let's do this for (15.7). Doing it for (15.8) is entirely similar. Consider the series expansion

$$\text{Arctan}\, x = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \cdots, \quad |x| < 1.$$

Now for $R > a$, the first term of (15.7) is

$$R\text{Arctan}\left(\frac{a}{R}\right) = R\left[\frac{a}{R} - \frac{a^3}{3R^3} + \frac{a^5}{5R^5} - \cdots\right] = a - \frac{a^3}{3R^2} + \frac{a^5}{5R^4} - \cdots,$$

and can be made arbitrarily close to a by choosing R sufficiently large.

The second term of (15.7) is

$$\frac{R}{\sqrt{R^2+b^2}} \cdot \frac{\sqrt{a^2-b^2}}{a} \cdot x - \frac{R}{(R^2+b^2)^{3/2}} \cdot \left(\frac{\sqrt{a^2-b^2}}{a} \cdot x\right)^3 + \cdots$$

For R sufficiently large,

$$\frac{R}{\sqrt{R^2+b^2}} \approx 1.$$

One can check that given x, the rest of the series can be made arbitrarily small by choosing R sufficiently large. Thus for any given x and all sufficiently large R, our spherical string length differs in magnitude from the ordinary distance function

$$a + \frac{\sqrt{a^2-b^2}}{a} \cdot x$$

by less than any preassigned $\epsilon > 0$.

Photons and Hyperbolas

Philosopher. Although we know that the sum-of-lengths property holds for spherical ellipses, my original question was about hyperbolas, not ellipses! Are we any closer to an answer?

Teacher. We're actually quite close, because a hyperbola is just an alternative projective view of an ellipse. Any ellipse on a hemisphere can be regarded as two antipodal ellipses on a sphere with opposite points identified. Now dehomogenize this sphere by selecting some other sub-hemisphere of the sphere; then if the boundary of the new hemisphere happens to cut each ellipse into two parts, we have a hyperbola on the hemisphere. Figure 15.5 shows an example.

Our cutting method of measuring on a hemisphere possesses a basic "invariance property" that lets us carry over string measurements to any new hemisphere. Here's the idea: on a hemisphere of radius R, a great semicircle has length πR, and two distinct points P and Q on that semicircle define two "travel lengths" L and $\pi R - L$. Now homogenize by extending the hemisphere to a sphere with two antipodal ellipses. After dehomogenizing at a different hemisphere, P and Q continue to define those same two lengths L and $\pi R - L$. That is, the pair of lengths is *invariant* with respect to the choice of hemisphere. Figure 15.6 suggests why:

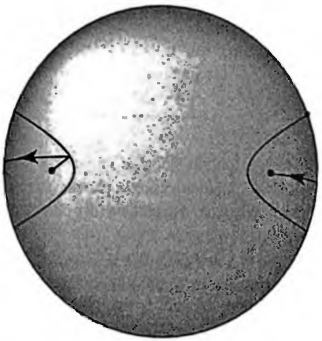


FIGURE 15.5. A photon released from the left focus bounces off the hyperbola, continues on a hemisphere geodesic, finally arriving at the right focus.

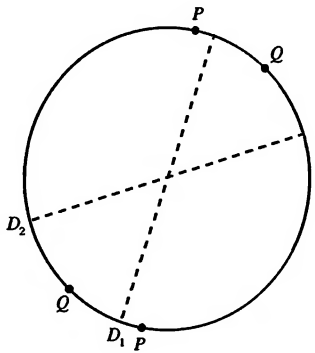


FIGURE 15.6. P and Q are two points on a sphere with opposite points identified, and the circle here is the sphere's great circle containing P and Q. The dashed diameters D_1 and D_2 divide the circle into semicircles in two different ways, corresponding to selecting hemispheres on the sphere in two different ways. Meditating a bit on this picture shows that no matter what diameter is chosen—that is, no matter what semicircle you work in—the "cutting method" of measuring on a hemisphere gives the same pair of long and short distances between P and Q on that semicircle. Our choices of D_1 and D_2 in this picture show that which is long and which is short can get reversed.

Interactive Applet

To see an arbitrarily rotatable version of Figure 15.5, see Chapter Applet 22 on the CD-ROM.

As we've done earlier, let's call a trip "long" if on that trip you cross the hemisphere boundary, and "short" otherwise. Then releasing a photon from an ellipse focus defines two short trips, while releasing one from a hyperbola focus defines one short and one long trip.

Philosopher. But after a photon strikes a hyperbola, if you force it to go backward, you get the usual wrong picture of two short trips, right?

Teacher. Yes. A photon "reflecting" off a conic means the photon doesn't cross the conic; reflecting off a hyperbola in that physically correct way gives one long and one short path. Actually, for real ellipses or hyperbolas, the foci can be thought of as defining the "inside" of that conic. Topologically, the conic divides $\mathbb{P}^2(\mathbb{R})$ into three parts—the conic itself, and two open regions. The region containing the foci we can call *inside*, and the other region, *outside*. [Compare Chapter 7, Exercise 10 on p. 118.] Physically, a photon released from a focus stays trapped inside the conic.

Student. It seems to me that the bigger R is, the *longer* the long path is!

Teacher. That's right.

Student. But isn't the limiting string constant $2a$?

Teacher. The traditional *wrong* one is! But **the string constant for physically correct reflection is the "diameter of the hyperbola."** By this, I mean the geodesic arc from one vertex to the other *that goes through the foci*, just as with ellipses:

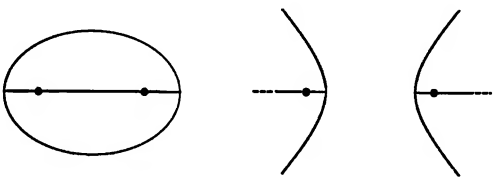


FIGURE 15.7. In each drawing, the diameter passes through the foci of a real ellipse or hyperbola.

So for the hemispherical hyperbola, the correct string length is not $R\text{Arctan}\left(\frac{a}{R}\right)$, but rather $\pi R - R\text{Arctan}\left(\frac{a}{R}\right)$. For large R, this is approximately $\pi R - 2a$.

Philosopher. This seems to help answer my question that started this whole discussion. In the usual plane hyperbola setup, when the photon bounces off the physically correct way, it indeed does start off on an unending journey, but we can regard the plane as a hemisphere with $R = \infty$. For any finite R, there's a corresponding finite string constant, $\pi R - R\text{Arctan}\left(\frac{a}{R}\right)$, which is very nearly $\pi R - 2a$ for R large.

Reality Check. We've been idealizing photons as little billiard balls that bounce off an ideally-reflecting conic. But what *really* happens? A photon travels a short way into the material of the mirror, meets an electron there and gets absorbed by it. After a short while the electron emits a *new* photon which travels in a direction statistically close to obeying the incidence-reflection property.

Teacher. Right. Of course $\pi R - 2a \to \infty$ as $R \to \infty$, but wherever we stand, for a finite but very large R, the hemispherical hyperbola is indistinguishable from the usual plane one—at least to our eyes. Suppose we release a photon from either focus, traveling in any direction. If we simply allow it to do the physically correct thing, it will eventually come back around, satisfying the very large string constant corresponding to R. In this way, using a very large hemisphere in place of a plane makes the sum-of-lengths property still hold.

Chapter Summary

In the stakes-and-string construction of an ellipse, the two string segments represent in a natural way paths of light rays before and after reflection. But the analogous, standard textbook picture for hyperbolas is physically wrong—the rays would have to pierce each branch to get to the other focus. We at last address this problem which was actually noticed at the very beginning of our journey. And it is a problem, because if we let a ray reflect in the physically correct way, one gets an infinitely long path. In that case, what does "sum-of-lengths property" mean? We adopt the point of view that a plane is a sphere of infinitely large radius. Then on a sphere of finite radius, we work out analogues of some basic plane measurements such as the Pythagorean theorem, location of conic foci, and so on. Importantly, we also extend to a sphere our signed string-length formulas in (5.5) and (5.6) [p. 77]. After setting up this machinery, we show that on a sphere (which can have a radius so large that locally it looks like a plane), the sum-of-lengths property *does* hold for hyperbolas when we let the light ray or billiard ball reflect in the physically correct way. These correct paths are the long ones, like flying west on a great circle from San Francisco to New York.

Exercises

1. Show that on a sphere of radius 5, a right triangle with legs 3 and 4 has hypotenuse approximately 4.79. Plot the hypotenuse length as a function of the sphere's radius. Note that by the time the radius has increased to 10, the hypotenuse is within 1% of the usual plane length of 5.0.
2. (a) Show that in Figure 15.2 on p. 288, the focal point on the sphere is the central projection of the point

$$\left(R \cdot \operatorname{Arctan} \sqrt{\frac{b^2 + R^2}{a^2 + R^2}},\ 0,\ R \right).$$

(b) Show that this last point can be alternatively written as

$$\left(R \cdot \text{Arctan}\sqrt{\frac{a^2-b^2}{b^2+R^2}},\ 0,\ R\right).$$

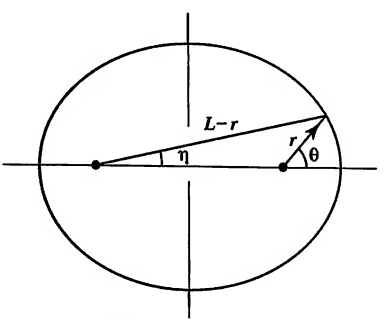
3. Figure 15.4 on p. 297 shows graphs only for $a = 3$ and $b = 2$. Keeping $b = 2$, how do the graphs change as you increase a beyond 3? What happens as a decreases to 2?

4. In Figure 15.1 on p. 286, suppose the right angle is at the north pole, and the triangle's sides extend to the equator. What does the spherical Pythagorean theorem say in this case? What if the triangle sides are both extended to the south pole? In general, what does the theorem say when the sides extend to the same circle of latitude?

5. Philosopher undoubtedly feels somewhat uncomfortable with that punch line in the box on p. 291. Why should the billiard ball decide upon the *long* axis? Why not the *short* one? What can you say to give him an "Aha!" experience?

6. **"Turning a Wave into a Particle"**: The box on p. 294 says that if you create a circular wave by simultaneously releasing photons from an ellipse focus, that wave, after repeated reflection, will look more and more like a particle moving along the major axis. *Why should this happen?* One can see the underlying reason by watching Chapter Applet 21. As the burst first begins, follow the photons on the left side of the circle; they start to draw relatively closer together after each reflection. After a while, the photons' original uniform spacing around the circle has changed to a dense, small arc centered about the major axis. Chapter Applet 20 is also helpful, because you can more precisely follow the history of a single particle through successive reflections, and you can also change the ellipse shape by varying the eccentricity. (If the eccentricity is close to 1, part of the ellipse goes off the screen, but diminishing the semi-latus rectum p will shrink the ellipse so you can see it all.) As a photon passes through a focus, its straight-line path forms an angle with the major axis, and the two applets seem to suggest that this angle becomes smaller with each succeeding reflection.

   To make this qualitative angle-diminishing observation quantitative, look at this diagram:



It appears that $\eta < \theta$. L is the major-axis length (that is, the string length), and r can be computed from the polar form of a conic in formula (A1.3), in which the origin is the right focus. Take $\phi = 0$ in this formula, so that the major axis is horizontal. Note that when $\theta \neq 0$ is small, the photon is heading almost directly *away* from the left focus.

(a) Redraw the above diagram, with the photon heading almost directly *toward* the left focus. To a casual onlooker, "the" angle from the horizontal might seem to increase after a reflection rather than decrease. Observe, using your diagram, that θ in fact decreases. (By changing the starting point and eccentricity, Chapter Applet 20 allows you to see many cases of this.)

(b) Show that

$$\eta = \text{Arcsin}\left(\frac{r\sin\theta}{L-r}\right),$$

where

$$r = \frac{p}{1+\varepsilon\cdot\cos\theta} \quad \text{and} \quad L = \frac{2p}{1-\varepsilon^2}.$$

[Hint: Observe that $\sin(\theta) = \sin(\pi - \theta)$; and use the law of sines.]

(c) Show that for $p \neq 0$, p doesn't affect the formula for η; thus we may assume without loss of generality that $p = 1$.

(d) Using Maple or Mathematica, graph

$$\eta = \text{Arcsin}\left(\frac{r\sin\theta}{L-r}\right)$$

for various choices of eccentricity $0 < \varepsilon < 1$. Observe the θ- and η-axis units, and notice that in all cases, $\eta \leq \theta$. Also, notice that when the ellipse is close to a circle (ε is small) the rate of angle decrease is slower; when the ellipse is cigar-shaped (ε close to 1), the decrease is faster.

(e) If we write

$$f(\theta) = \text{Arcsin}\left(\frac{r\sin\theta}{L-r}\right),$$

then for any fixed choice of eccentricity $0 < \varepsilon < 1$, successive angles of reflection are given by successive compositions of f with itself. Graph the functions $f(f(\theta))$, $f(f(f(\theta)))$, and $f(f(f(f(\theta))))$. Explain why these graphs suggest that for any particular ellipse, the photons converge toward moving along the major axis.

(f) Using the Maclauren series for sine, show that for small θ,

$$\eta \approx \left(\frac{r}{L-r}\right)\theta.$$

Next, for any given ellipse, find a constant c, so that

$$\frac{r}{L-r} < c < 1$$

holds for all small θ. Now try to prove that the photons in fact converge to motion along the major axis.

7. Simultaneously release 100 photons from the focus of a parabola. As they begin their journey, assume they are equally spaced on the expanding circle.
   (a) After they all reflect, what is the shape of the moving "wall" of photons?
   (b) In this wall of 100 photons, what is the distance D_i of each photon from the x-axis?
   (c) To get an idea of the reflected photon "density," plot $\frac{1}{D_i}$.

16

How Conics Solved a 2000-Year-Old Problem

in which an understanding of heavenly-body orbits finally came down to earth

Lost in the Mists of Time

Philosopher. It seems amazing to me that although conics go back more than 2300 years, in our explorations we've nonetheless managed to add to our understanding of them. We've done it mostly by looking in their natural habitat, complex projective space. We've peeked into the fourth dimension, seeing a pair of foci on each principal axis. And to each pair of foci, we've discovered an associated pair of directrices. We have computed the area of hyperbolas and found the everywhere-negative surface curvature of conics. And so on. It's been especially gratifying to see what a reliable guide *symmetry* can be, so often leading us to a deeper and more complete understanding!

Student. Over 23 centuries—that's really *old*! Where did conics actually come from?

Teacher. We can only make educated guesses. Reliable clues to the origin of conics have been lost in the mists of time. Geometers working around 500 BCE tried to solve a wide variety of geometric construction problems using only straightedge and compass. But in particular, there were three problems—doubling the cube, squaring the circle and trisecting arbitrary angles—that seemed to defy their most clever and persistent attempts. It is thought that conics were devised as a way to help solve these problems.

Student. Exactly what are those problems?

Teacher. "Doubling the cube" means constructing the edge of a cube having double the volume of a given one. If the original cube has edge a, the new one has edge $\sqrt[3]{2}a$, so this problem really comes down to constructing $\sqrt[3]{2}$.

"Squaring the circle" means constructing a square having the same area as a given circle. If the original circle has area πr^2, then the square has side $\sqrt{\pi}r$, so this problem essentially comes down to constructing a line segment of length $\sqrt{\pi}$. Of course in any numerical construction problem, it's always assumed that there's a *unit length*, which is taken to be the length of some line segment, singled out in that construction for that purpose. Then the number is to be constructed in terms of that unit segment, using straightedge and compass.

Student. And I know about trisecting an angle. That's like bisecting an arbitrary angle, except the goal is to divide it into three equal parts instead of just two.

Teacher. Yes. The important word here is *arbitrary*, since there are certain special angles, like a straight or right angle, that can be trisected.

It has been proved that none of these three construction problems can be carried out in finitely many steps using only straightedge and compass. In desperation—and desperation has proved to be the germ of a surprising number of advances in science—the ancients attempted to expand their toolbox to help solve such problems. Conics became part of that toolbox.

The list of mathematicians who found approximate solutions to the cube-duplication problem is impressive. Here are some ancient contributors, with approximate dates:

- Hippocrates (440 BCE)
- Archytas (400 BCE)
- Eudoxus (370 BCE)
- Menaechmus (350 BCE)
- Eratosthenes (230 BCE)
- Nicomedes (230 BCE)
- Apollonius (225 BCE)
- Diocles (180 BCE)

In the 17th century, Descartes, Huygens and Newton all became interested in the problem and constructed approximations.

Student. How can conics have anything to do with these problems?

Teacher. Take the cube-doubling problem, for example, where the task is to construct $\sqrt[3]{2}$. Suppose you could construct the two parabolas $x = y^2$ and $2y = x^2$. It's then easy to see that one point of intersection is $(2^{2/3}, 2^{1/3})$. To see why, notice that $x = y^2 \Rightarrow x^2 = y^4$; substituting this last into $2y = x^2$ to eliminate x gives $y^4 = 2y$, or $y^3 = 2$. Of course geometers in the 5th century BCE didn't have a rectangular coordinate system as we know it today. That was developed from early forms invented by Fermat and independently, Descartes. Nonetheless, the intersection point $(2^{2/3}, 2^{1/3})$ in Figure 16.1 is a corner of the indicated rectangle with sides $2^{2/3}$ and $2^{1/3}$.

Philosopher. Using a straightedge and compass, they could construct parabolas?

Teacher. They could construct any finite number of points on certain parabolas, thus getting arbitrarily good approximations of those parabolas by connecting successive constructed points. For example, the upper half of the parabola $x = y^2$ can be looked at as the "square-root curve" $y = \sqrt{x}$, and the ancients knew very well how to construct square roots. In fact more generally, given lengths a and b, they could construct the geometric mean $\sqrt{ab}$, as Figure 16.2 suggests.

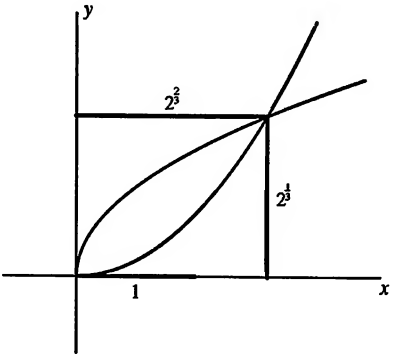


FIGURE 16.1. The parabolas $x = y^2$ and $2y = x^2$ intersect at $(2^{2/3}, 2^{1/3})$—the upper right corner of the rectangle. Once you have this point, you've got the vertical side length of the rectangle, which is $2^{1/3}$.

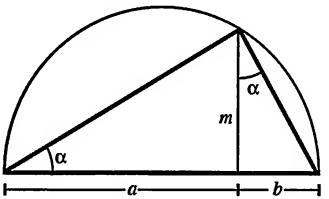


FIGURE 16.2. The altitude m of the heavily drawn triangle is the geometric mean of lengths a and b. This is because the two smaller triangles are similar, meaning that $\frac{a}{m} = \frac{m}{b}$, which implies $m^2 = ab$—that is, $m = \sqrt{ab}$. Choosing $b = 1$ allows one to find square roots: the altitude m is then $\sqrt{a}$.

Constructing a succession of semicircles as in Figure 16.3 thus produces square roots (vertical lengths) of a succession of x-values (horizontal lengths). Connecting the tops of these triangle altitudes gives an approximation to the curve $y = \sqrt{x}$.

Student. Looks like you've already got one of the curves you need!

Interactive Applet

For an animated, interactive version of Figure 16.3, see Chapter Applet 23 on the CD-ROM.

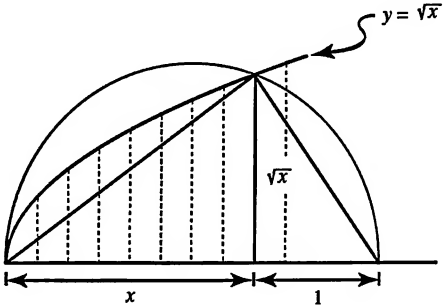


FIGURE 16.3. The square root construction can be applied repeatedly to get points on the square-root curve, $y = \sqrt{x}$, shown here. Connecting these points creates the approximation to this curve constructed by the ancient Greek geometers.

Teacher. Right. The other curve lies on a parabola opening vertically rather than horizontally—in modern notation, $x = \sqrt{2y}$. Regarding $\sqrt{2y}$ as the geometric mean of 2 and y amounts to constructing the curve via a succession of geometric means of 2 and y along a vertical line:

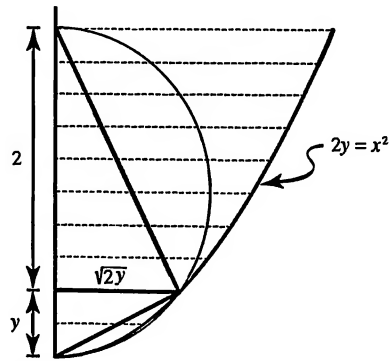


FIGURE 16.4. In this illustration, the right half of the parabola $2y = x^2$ is approximated by constructing a succession of geometric means of 2 and y. Each horizontal line segment is an altitude of a right triangle inscribed in a circle of diameter $2 + y$.

Doing everything in one diagram produces the two intersecting parabolas of Figure 16.1, which gives a line segment of length $\sqrt[3]{2}$.

Philosopher. Those ancients were clever, weren't they? It's easy to forget that they were every bit as human, determined, persistent and creative as we are. I guess one of the great things about civilization is that wisdom of one generation gets handed down to succeeding ones. Today, thanks to the successes and failures of the ancients and others before us, we are more attuned to the importance and power of symmetry. Just as we've been doing here, other mathematicians and physicists today apply symmetry to make progress.

Student. But a lot of thinkers a long time ago made a really big thing out of symmetry. I read once how the ancients believed the heavenly bodies were perfect spheres. And Plato (428?–347 BCE) thought the planets had to move in perfect circles. And Kepler got pretty excited about those nested platonic solids, didn't he?

Teacher. You raise some very salient points. I guess maybe we should say that today, we are more *skilled* in applying symmetry. We've learned that trying to apply symmetry to large-scale objects—such as determining the shape of heavenly bodies, the shape of the universe, and so on—is generally inappropriate. At the very smallest scales—quantum and smaller, symmetry has had far greater success. The smaller, the more "ultimate" you get, the more often symmetry turns out to be helpful. The big push toward string theory is heavily based on symmetry and what physicists call "supersymmetry," a more pervasive symmetry. And in pure mathematics, symmetry can be a very potent force. In our odyssey, we've been lucky enough to catch a symmetry wave and surf on it!

Student. Anyway, what about the origins of conics?

Teacher. The initial success these ancients had with conics were followed by detailed investigations of them. Apollonius, generally regarded as the greatest ancient mathematician after Archimedes, wrote a monumental work, *Conics*. It consisted of eight books, of which seven survive. In them, he worked out hundreds of facts about conics. Never since in the ancient world of mathematics was any work produced of such quality, subtlety, and completeness.

Student. Wow—paper over 2000 years old! I'll bet it's pretty fragile by now.

Teacher. The original was probably written on rolls of papyrus, not on paper, but in any case we can be quite sure that all traces of it have been lost. Through history, scribes in monasteries copied what appeared to be important documents, giving them a succession of new leases on life. Both we and Renaissance thinkers owe a tremendous debt of gratitude to those legions of nameless scribes whose work kept alive the glowing embers of knowledge and wisdom. And in the ancient world itself, with Gutenberg's printing press far in the future, scribes also functioned as publishing houses for the dissemination of written material.

Philosopher. It was Apollonius who added a second nappe to a cone. Pre-Apollonian hyperbolas had only one branch.

Teacher. That's true—the great geometer already had his fingers on the pulse of symmetry, didn't he?

Philosopher. Without the second half of a cone, we'd miss how hyperbolas form closed loops—we'd never see the geometric unity between them and ellipses.

The terms *nappe*, *napkin*, *(n)apron* and *map* are all related, deriving from the original Latin *mappa*, meaning "tablecloth." One can think of a tablecloth as a large sheet, while an apron or napkin are smaller ones. A wall map is drawn or printed on a large sheet of paper. A double-napped cone is a two-sheeted cone, much like a hyperboloid of two sheets. In fact, $x^2+y^2-z^2=r^2$ describes such a hyperboloid, and as $r \to 0$, it morphs into the usual double-napped cone. Visually, $r \to 0$ corresponds to viewing the hyperboloid from an ever-greater distance; to our eyes, the two hyperboloid "noses" approach each other and become more pointed.

Student. But where did the cone itself come from? I mean, I don't see any cones in the construction of those parabolas in Figures 16.3 and 16.4.

Teacher. That's a very good question, and we really don't know the answer. A man of Apollonius' cleverness may have looked at a succession of circles—analogous to the semicircles above—as level curves of a cone in three dimensions. Of course level curves are plane sections. But this is just an educated guess.

How Move the Wanderers?

Teacher. Although a handful of ancient thinkers wondered and dreamt about conics, a problem of more widespread and immediate concern back then was understanding how heavenly bodies move—after all, they were there for all to see and wonder about. Just what paths do they take? The "wanderers"—the planets—posed a special problem, because although they often seemed to move reasonably enough, sometimes a planet would appear to slow down, and then go *backwards* for a while, before resuming its normal direction of travel. This so-called "retrograde motion" fascinated, perplexed, and sometimes frightened.

Philosopher. As we've noted, Plato assumed that planetary orbits had to be based on the circle, which he regarded as the most perfect of all curves. Apollonius himself, around 200 BCE, had the idea of describing planetary motions by combinations of circular motions. From Ptolemy through much of the Renaissance, geometers took these ideas very seriously. Now there's a thin line between persistence and obstinence, and when simple combinations of circles seemed to offer no simple solution to planetary motion, they became all but blinded by an obsession to "save Plato's circles." Constructions became ever more complex, and their obsession ended up inspiring a lot of mathematics. For example, they found that by appropriately rolling circles on other circles, they could create various sorts of retrograde motion, and they became experts on cycloids of many varieties, creating models of the solar system that would have impressed even Rube Goldberg.

We know now that retrograde planetary motion is *perceived* motion. It arises when the observer views the heavens on a platform that *moves*! Church dogma of the day decreed that the platform—our earth—didn't move. It was the center of the universe, fixed for all time. Nicolaus Copernicus (1473–1543) made careful measurements of planetary motions, and eventually concluded that the planets had to move around the sun, not the earth. This brilliant thinker, justifiably wary of retribution, cautiously distributed a very limited number of "handbooks" he'd made, written in Latin so as to minimize the likelihood of stirring

Retrograde motion constantly takes place on the wheels of a moving train. To keep the wheels from sliding off the track, each wheel has a flange dipping below the track's top surface. If you paint a small white dot on the boundary of a wheel, then whenever the dot is below the track surface, it moves backward as the train moves forward. The curve traced out by the dot is called a "prolate cycloid."

up the less literate masses. But he waited until near the end of his life to publish his masterpiece, clearly outlining experimentally verified reasons why the nearby universe must be sun-centered. He saw the first published copy on his deathbed.

Student. This is amazing!

Teacher. A generation later, Galileo Galilei—also brilliant, but as brazen as he was eloquent—eventually concluded that Copernicus was right, that our system of planets, sun, and moon was indeed sun-centered. His forthright talk about this led to paroxysms in the church, which felt that such thinking was extremely dangerous to ordinary mortals. But Galileo was not an ordinary mortal; he was a celebrity who attracted students from hundreds of miles away to attend his inspired lectures in Pisa. And ordinary folk from the general populace would clamor for a peek through his state-of-the-art telescope. For a time, this protected him from the full wrath of the church. But toward the end of his life he was forced to recant, and was sentenced to house arrest.

Gallleo had enemies, and his problems with the church stemmed from more than just honest science. In his early years of teaching, he devoted considerable time, effort and hard-earned money in setting up a home laboratory and running experiments to test the "physics laws" of the day. He just couldn't stand it when his fellow teachers unquestioningly chanted physical dogma—often false—handed down from centuries earlier. The young, brash Galileo made the mistake of calling them "fat old watermelons gone to seed." They may not have been good scientists, but they had long memories, and many years later, after Galileo had written his famous *Dialogue Concerning the Two Chief World Systems*, his enemies convinced the church authorities that Galileo intended *Simplicio* (the "simpleton" among the *Dialogue*'s three characters) to mean the Pope. Any good feelings the Pope might have had for Galileo dried up, and Galileo's troubles took on more serious proportions. He faced the dreaded Inquisition and was found guilty of heresy. The great icon was sentenced to house arrest for the remainder of his life.

During those last years he worked on what he considered his greatest work, which today we might call "terrestrial physics." He showed, for example, that a cannonball follows a parabolic arc. He was on to conics.

Student. Yes, yes! I had to derive the cannonball path once on a test, and I got it perfectly right!

Teacher. Around this time, the much quieter and more reclusive Kepler was doggedly seeking a mathematically correct description of planetary motion—we might call it "celestial physics." It was a decades-long quest, and realizing this dream proved to be torturously difficult. It meant extracting reliable astronomical data from the arrogant, wealthy and egocentric Tycho Brahe; and, after Tycho's death, further coaxing and begging from his family members who jealously guarded the most precise astronomical data ever collected up to that time.

> Were it not for Tycho Brahe (1546–1601), astronomy and our understanding of the heavens would probably have stagnated for many more decades. He demonstrated an early, intense interest in celestial bodies, and observed them like a true professional. (He's usually called by his first name, pronounced "*Tee*-Ko.") He caught the attention of the Danish king, Frederick II, who in 1575 gave him the small island Hveen, some fourteen miles north of Copenhagen, as well as generous funds with which to build an observatory, replete with staff—of which Kepler was eventually a member. This was all before telescopes. Brahe built a "naked eye" observatory second to none. Its largest quadrant had a radius of 31.5 feet, was built into a very sturdy wall, and could measure a star's elevation above the horizon with an accuracy of 1 minute of arc—about the limit of human vision. Before him, astronomers measured only special events, as when a planet went retrograde. But Tycho made many, many measurements spaced along a planet's path, even averaging multiple readings for increased accuracy. Tycho also carefully analyzed the comet of 1577, traveling all over Europe to take measurements from various locations, concluding from parallax considerations that the comet was far beyond the orbit of the moon. This contradicted Aristotle, who taught that meteors and comets arose from the atmosphere. The word "meteorology" arose from this false notion.

In his lonely quest that lasted many years, Kepler performed reams upon reams of calculations. It took decades before he at last gave up—with the greatest of reluctance—on "saving the circles." It proved to be the *ellipse*—to Kepler a disturbingly imperfect circle—which precisely modeled Tycho's data on Mercury. His conclusions eventually became distilled into "Kepler's Three Laws of Planetary Motion":

I Each planet moves in an elliptical orbit with the sun at one focus;

II The line segment from the sun to any particular planet sweeps out equal areas in equal times;

III The period of any planet is proportional to $a^{3/2}$, where a is the semi-major axis of the planet's elliptical orbit.

He had elucidated the paths of the planets as no one before him, with simplicity and precision. He'd cracked a centuries-old problem.

Student. He must have been tremendously proud of that!

Teacher. Actually, it was very hard for him to square his deep religious convictions with what he'd discovered. He rationalized what seemed to him a sad state of affairs by reasoning that if God made man imperfect, then why shouldn't He have made physical laws imperfect, too?

> In modern terms, we could summarize Kepler's first law as the solution to a "curve fitting problem"—he found a second-degree curve that best fit (most nearly passes through) the data-points gathered by Tycho.

The Ellipse Gets a New Equation

Student. This story is fascinating. But the sun being at a focus bothers me. It seems reasonable enough to place the sun at the origin, but then when you write the equation for the planet's ellipse, if the sun is actually at a focus, meaning the focus is at the origin, that should make the equation a lot more complicated.

Philosopher. That bothers me, too! After all, the equation

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}=1$$

is so nice. Wherever the sun is, it seems natural to put the ellipse's *center of symmetry* at the origin, rather than putting a focus there. Moving the ellipse elsewhere, though certainly possible, has got to make things more complicated.

Teacher. Maybe we're looking at the rectangular coordinate system as the be-all and end-all, here. Usually it's a wonderful coordinate system, but actually, rectangular coordinates are rather arbitrary when it comes to planetary motion. The big gravitational force here is from the sun, and no matter where or how a planet may move, the gravitational force vector lies along the line through the sun and planet. If we think of the sun as stationary, then the force it exerts on a planet can be represented by a vector aimed from the planet toward the sun's fixed location. As the planet moves, so does the vector.

Philosopher. That sounds like polar coordinates!

Teacher. That's right. So rather than writing a rectangular-coordinate equation of an ellipse having a focus at the origin, let's try writing it in polar coordinates.

Philosopher. How would you do that?

Teacher. Start by thinking of a planet going around the sun, duly obeying Kepler's first law. Suppose the ellipse has diameter $2a$ and eccentricity ε, and look at the triangle in Figure 16.5. We can express each side of this triangle in terms of r and a: by sum-of-lengths, one side has length r and another, $2a-r$; the third, horizontal side has length $2F$, but since the eccentricity is $\varepsilon=\frac{F}{a}$, this side is $2\varepsilon a$. These three lengths, r, $2a-r$ and $2\varepsilon a$

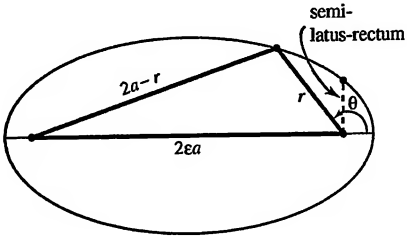


FIGURE 16.5. The law of cosines leads from the sum-of-lengths property of an ellipse to a polar equation for it. The triangle has one vertex on the ellipse and the other two at the foci.

are related through the law of cosines:

$$(2a - r)^2 = r^2 + (2\varepsilon a)^2 - 2r(2\varepsilon a)\cos(\pi - \theta).$$

Expanding this gives

$$4a^2 - 4ar + r^2 = r^2 + 4\varepsilon^2 a^2 + 4r\varepsilon a \cdot \cos\theta.$$

Simplifying and collecting terms leads to

$$a(1 - \varepsilon^2) = r(1 + \varepsilon \cdot \cos\theta).$$

That is,

$$r = \frac{a(1 - \varepsilon^2)}{1 + \varepsilon \cdot \cos\theta}.$$

The parameter a gives the size of the ellipse, and ε gives its shape. It is easy to show that the numerator $a(1 - \varepsilon^2)$ is the length of the "semi-latus-rectum"—depicted traditionally in Figure 16.5 as the vertical line segment from a focus to the ellipse. Notice that when $\theta = 90°$, $\cos\theta = 0$, so

$$r = \frac{a(1 - \varepsilon^2)}{1}.$$

Thus $\theta = 90°$ gives the dashed semi-latus-rectum of Figure 16.5. Its length is denoted by p, so our equation becomes

$$r = \frac{p}{1 + \varepsilon \cdot \cos\theta}. \tag{16.1}$$

So a and p represent two different ways of controlling ellipse size—a specifies the semi-major axis, while p specifies the semi-latus rectum.

These polar forms have a very appealing property: to rotate the ellipse by ϕ, we merely follow our earlier action-reaction philosophy, subtracting ϕ from θ. This gives

$$r = \frac{p}{1 + \varepsilon \cos(\theta - \phi)}. \tag{16.2}$$

Now compare (16.2), which has an immediately-obvious size, shape and tilt, with the equation of the tilted conic, centered at the origin, in rectangular coordinates:

$$Ax^2 + Bxy + Cy^2 = 1.$$

Can you immediately tell me the size, shape and tilt of *that*?

Philosopher. No, not at all. Sometimes you can be quite persuasive.

Teacher. An ellipse, written in polar coordinates and with a focus at the origin, has a big geometric advantage!

Philosopher. From your derivation, it seems that the origin is at the right focus of the ellipse.

Teacher. Yes. We could have placed the angle θ at the left focus, still working with the triangle in Figure 16.5. Then instead of $\cos(\pi - \theta) = -\cos\theta$, we would have just $\cos\theta$. The left-focus version of (16.2) is then

$$r = \frac{p}{1 - \varepsilon \cos(\theta - \phi)}. \tag{16.3}$$

Student. That's only an ellipse. What about a parabola or a hyperbola?

Teacher. Though we derived formulas (16.1)–(16.3) for an ellipse, one can check that they can also define a parabola or hyperbola—it's merely that the eccentricities are different. The case of a parabola is interesting, since one of the foci is at infinity.

Student. Doesn't that mean (16.2) and (16.3) give the same parabola?

Teacher. No—for simplicity, assume $\phi = 0$. Then you can think of a "right" or "left" focus as being at infinity, and that's the side on which the parabola opens. For example, the parabola in (16.1), has its origin at the "right" focus, the other focus being at infinity; in this case, the parabola opens up to the left. With $\varepsilon = 1$ and $\phi = 0$, (16.3) defines a parabola opening to the right.

Student. If I take $\varepsilon = 1$ and $\phi = \frac{\pi}{4}$ in (16.3), I'll get a parabola tilted 45°?

Teacher. Exactly.

> Because ε is the shape parameter, and because $\varepsilon = 1$ for a parabola, all parabolas in fact have the same shape—in terms of Euclidian geometry, all parabolas are similar. This is often not appreciated, because when we zoom in toward the vertex of a parabola, what we see on a piece of paper or a monitor looks less curved, while zooming away, it looks sharper at the vertex and longer overall. In fact, from a very long way off, a parabola looks to our eyes like a ray (half line).

Philosopher. So for the same ϕ, (16.2) and (16.3) give parabolas pointing in opposite directions?

Teacher. Yes, and we can also see this by adding π to the tilt ϕ, which interchanges (16.2) and (16.3).

Enter Newton

Student. I must be missing something. Kepler's Laws seem really good, but suppose I was looking through a telescope and just happened to notice a big asteroid heading our way! Wanting to save the world, I at once report my discovery to a couple of astronomers. How are they going to use Kepler's Laws to figure out whether or not the asteroid will hit us? It seems to me that the faster we find the answer, the better!

Teacher. You've asked a very basic question. Kepler's Laws describe properties of a planetary path as a *fait accompli*. In truth, however, that completed curve is the end result of a succession of instantaneous changes. Even ignoring the gravitational attraction on a given planet from all matter in the universe except for the sun, these instantaneous changes are not trivial. For one thing the force on the planet from the sun is constantly varying, and that continually changes the planet's speed. Of course, the distance from the sun is always varying, too. For some heavenly bodies—possibly including your asteroid—such changes can mean the body moves not along an ellipse, but rather along a parabola or hyperbola. Kepler's Laws don't mention parabolas or hyperbolas. What is needed is a method encompassing the more general paths that the asteroid or planet may be following—some way of using "initial data" (position and velocity at some instant) to determine the path. It's here that Newton enters the picture.

Student. Are you saying that what he did will answer my question?

Teacher. Yes.

Student. Tell me!

Teacher. Let me begin with a bit of philosophy. You are doubtless aware that breaking things up into smaller, more fundamental building blocks can frequently lead to conceptual simplification, unity and progress. Just think how discovering the chemical elements transformed the conceptual framework of chemistry! Perhaps even more remarkably, think how just three elements—C, H and O (carbon, hydrogen and oxygen)—form the chemical basis of organic chemistry and all of life.

Student. Sounds like you're building up to something big.

Teacher. Maybe I'm stretching an analogy a bit, but what Kepler did was a little like discovering not chemical *elements*, but only various *compounds*. He identified entire curves—ellipses—as the paths of planets. This is surely important, but this discovery doesn't tell us how these curves are formed. There's inherently something that produces the planetary paths, and Kepler's discovery says nothing about that. What Newton did was somewhat analogous to breaking up a compound into its constituent atoms. He broke up an entire curve into "building blocks" and then figured out how its tiny parts were put together.

Student. What on earth are "building blocks of a curve"?

Teacher. Think of a curve as a concatenation of many tiny arcs. The whole curve itself is the result of appropriately stringing these together.

Student. That sounds like something I once saw when I used some software to solve differential equations. I saw the solution curve being drawn on the screen, almost like an ant dragging a miniature paintbrush—it was exciting.

Teacher. I believe that software was computing a succession of differentials of arc and stringing them together—plotting them—on the screen. Newton's challenge was to find just how these little arcs were strung together, a bit like a chemist figuring out how his atoms are connected. Take a little arc so small, that there's only a negligible difference between it and the line segment with the same endpoints. You can think of a "differential" as a kind of process in which the length of these approximating line segments tends to zero. You can let this process continue not only to the stage where you can no longer visually distinguish between the curve and the union of the minuscule line segments, but to the stage where even any difference in numerical results approaches the limit of your computing device.

Student. What do you mean by "how the little pieces are put together"? Tiny arcs or line segments are joined one after the other, aren't they?

Teacher. Well, there's the issue of angles. Your on-screen differential equation solver not only drew tiny line segments, but also decided on the correct slope of each succeeding segment.

Student. Like a whole pile of Tinker Toys, together with instructions for taking the wooden sticks—differentials—and inserting them into the holes in those wooden wheels to create the correct angles of the desired object!

Philosopher. An interesting analogy! Mathematically, Newton's eventually described how the little arcs were strung by formulating a "differential equation." The Tinker Toy as-

Sir Christopher Wren, Robert Hooke and Sir Edmond Halley struggled with the problem of what force-law would cause planets to move along Kepler's ellipses. Hooke originally thought the force between the sun and any planet was "linearly increasing," as if the ends of an enormous but invisible spring were attached to the sun and planet. This erroneous idea eventually evolved into suspecting an inverse-square law, but Hooke was unable to determine what kind of orbit such a law would make the planets move in. But Edmond Halley knew just the person to ask. Newton's answer was immediate: "Why, in ellipses, of course." The astonished Halley wanted to know how he knew that. Newton replied that he'd worked out the answer some twenty years before, but never published it. Ironically, part of the reason Newton didn't publish was Hooke's (perhaps jealous) criticism of his theory of light, years earlier. The thin-skinned Newton swore he'd never publish again. Not only did Halley recognize the incredible importance of Newton's work, he was greatly supportive of it. He patiently worked on a reluctant Newton, finally convincing him to write up his results. After Newton's masterpiece *Principia* was completed, Halley totally financed its publication.

sembly process is like solving a sort of "poor man's differential equation." Stick-and-ball chemical model building would be another example.

Teacher. I agree. After forming an appropriate differential equation, the next thing is to solve it. The evolving curve on the screen represents one way of solving it: numerically. That's something akin to the parabola-construction of the ancient Greeks, in that the solution can be approximated as closely as we want. But it is still only approximate. For the idealized case of a point-mass planet influenced by a fixed star, the path is a conic, and for this, Newton wanted the exact equation of the conic. He succeeded.

Student. What a genius!

Newton's Great Differential Equation

Teacher. Using modern terminology and notation, let's retrace Newton's basic idea. We phrase the overall problem this way:

> **Let a fixed point of unit mass M create a gravitational field, and let another point having mass m move due only to the influence of this field. If m is much smaller than M, and if m's position and velocity are known at some instant t, then determine the equation of m's path.**

By assuming that m is much smaller than M, our assumption about M being fixed becomes very nearly true. Now our problem is actually a *two*-parter: finding the differential equation; then solving it.

We now take up the first part, finding the equation. This turns out to be a little symphonette based on the chain rule of differentiation.

We use polar coordinates throughout, with the pole at the origin. As m moves, its polar coordinates $(r(t), \theta(t))$ of course vary with time. For brevity, we often write these time-dependent coordinates as simply (r, θ). As for the force between the two masses, we assume Newton's universal gravitational attraction law

$$F = G\frac{mM}{r^2},$$

with $M = G = 1$, making the force

$$F = \frac{m}{r^2}.$$

This force on m is always directed toward the origin. M therefore acts as a "center," and the force in this case can be descriptively called a *central force*. There are two moving orthogonal unit vectors based at m which are natural to polar coordinates. These are shown in Figure 16.6, denoted by $\mathbf{I}(t)$ and $\mathbf{J}(t)$.

Sometimes we write simply $\mathbf{I}$ and $\mathbf{J}$. The force vector on m at any time can be expressed as

$$\mathbf{F}(t) = -\frac{m}{r^2} \cdot \mathbf{I}(t). \tag{16.4}$$

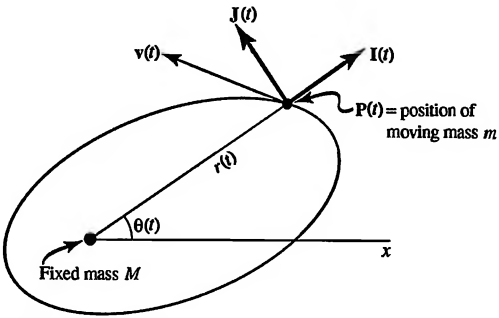


FIGURE 16.6. Think of M as a stationary sun at a point which will turn out to be the focus of an ellipse. $\mathbf{P}(t)$ is the position of a planet orbiting the sun. $\mathbf{I}(t)$ and $\mathbf{J}(t)$ are moving unit vectors based at $\mathbf{P}(t)$, with $\mathbf{I}(t)$ always pointing directly away from M.

Of course the motion of m is determined from "$F = ma$" which can be written more completely and informatively as a *vector* equation

$$\mathbf{F}(t) = m\mathbf{a}(t). \tag{16.5}$$

Since acceleration $\mathbf{a}$ is the second derivative of position with respect to time, (16.5) is indeed a (vector-valued) differential equation. Now from (16.4), we know the left side of (16.5): it's

$$-\frac{m}{r^2} \cdot \mathbf{I}(t).$$

What about the right side of (16.5)? Because $\mathbf{F}$ is directed toward the origin, so is $\mathbf{a}$. The question then becomes one of finding $\mathbf{a}(t)$ in terms of the position. In terms of the natural basis $\{\mathbf{I}(t), \mathbf{J}(t)\}$, the position vector is

$$\mathbf{P}(t) = r(t)\mathbf{I}(t). \tag{16.6}$$

We need to differentiate this twice, and since this is a product of two varying quantities, we use the product rule.

Here's the first differentiation with respect to t:

$$\begin{aligned}
\mathbf{v}(t) = \dot{\mathbf{P}}(t) &= \frac{d[r(t)\mathbf{I}(t)]}{dt} \\
&= \dot{r}(t)\mathbf{I}(t) + r(t)\dot{\mathbf{I}}(t) \qquad \text{(product rule)} \\
&= \dot{r}(t)\mathbf{I}(t) + r(t)\frac{d\mathbf{I}}{d\theta} \cdot \frac{d\theta}{dt} \qquad \text{(chain rule)} \\
&= \dot{r}(t)\mathbf{I}(t) + r(t)\mathbf{J} \cdot \dot{\theta}.
\end{aligned}$$

In writing this last line, we've used

$$\frac{d\mathbf{I}}{d\theta} = \mathbf{J}.$$

Student. Is that obvious?

Teacher. Well, $\mathbf{I}$ and $\mathbf{J}$ are both *unit* vectors, so their tips are on a unit circle whose center is at $\mathbf{P}$. When θ increases by $\Delta\theta$, the $\mathbf{I}$-vector moves to a nearby point on the circle. In obvious notation, $\mathbf{I}(\theta + \Delta\theta) - \mathbf{I}(\theta)$ is a "secant vector" connecting these two points, and its length is well-approximated by the arc length subtended by $\Delta\theta$. Since the circle has radius 1, this arc length is of course $1 \cdot \Delta\theta$. One can now easily complete the argument that

$$\lim_{\Delta\theta \to 0} \left(\frac{\mathbf{I}(\theta + \Delta\theta) - \mathbf{I}(\theta)}{\Delta\theta} \right)$$

is tangent to the circle, points in the direction of $\mathbf{J}$, and has unit length. In short, the limit *is* $\mathbf{J}$—that is to say,

$$\frac{d\mathbf{I}}{d\theta} = \mathbf{J}.$$

Student. It seems that by using the same kind of reasoning, one could show that

$$\frac{d\mathbf{J}}{d\theta} = -\mathbf{I}.$$

Teacher. That's exactly right.

Continuing, let's now differentiate $\mathbf{v}(t)$ with respect to t:

$$\begin{aligned}\mathbf{a}(t) = \dot{\mathbf{v}}(t) &= \frac{d[\dot{r}(t)\mathbf{I}(t) + r(t)\dot{\theta}\mathbf{J}]}{dt} \\ &= \left[\ddot{r}(t)\mathbf{I}(t) + \dot{r}(t)\dot{\mathbf{I}}(t)\right] + \left[\dot{r}(t)\dot{\theta}\mathbf{J} + r(t)\ddot{\theta}\mathbf{J} + r(t)\dot{\theta}\dot{\mathbf{J}}\right].\end{aligned}$$

This last line involves the product rule applied to two and to three factors. In the next step, we use the chain rule:

$$\begin{aligned}&= \left[\ddot{r}(t)\mathbf{I}(t) + \dot{r}(t)\frac{d\mathbf{I}}{d\theta} \cdot \dot{\theta}\right] + \left[\dot{r}(t)\dot{\theta}\mathbf{J} + r(t)\ddot{\theta}\mathbf{J} + r(t)\dot{\theta}\frac{d\mathbf{J}}{d\theta}\dot{\theta}\right] \\ &= \ddot{r}(t)\mathbf{I}(t) + \dot{r}(t)\mathbf{J} \cdot \dot{\theta} + \dot{r}(t)\dot{\theta}\mathbf{J} + r(t)\ddot{\theta}\mathbf{J} - r(t)\dot{\theta}\mathbf{I}(t)\dot{\theta}.\end{aligned}$$

This final step uses what we observed just a moment ago, that

$$\frac{d\mathbf{J}}{d\theta} = -\mathbf{I},$$

and from this final expression we can pick off the two $\mathbf{I}$-terms—namely,

$$[\ddot{r}(t) - r(t)(\dot{\theta}(t))^2]\mathbf{I}(t).$$

Student. Why are you picking the $\mathbf{I}$-terms?

Teacher. Because $\mathbf{I}$ is the only direction in which there's actually any gravitational attraction—that force is *central*, lying in the direction of $\mathbf{I}$.

So at this stage, we know that since the vector equation $\mathbf{F}(t) = m\mathbf{a}(t)$ involves only the I-component, then although this component moves with time, the strength-of-force equation for our central gravitational field, is

$$\frac{m}{r^2} = m(\ddot{r} - r\dot{\theta}^2)$$

—that is,

$$\frac{1}{r^2} = \ddot{r} - r\dot{\theta}^2. \tag{16.7}$$

Philosopher. I thought you said the differential equation might be easy to solve! Certainly no solution leaps out at me.

Sir Edmond Halley
1656–1742

A "scientist for all seasons," Halley was an astronomer of major importance, is considered the founder of geophysics, was a pioneer in social statistics and an expert on physical instrumentation. He noticed the close similarity in orbits of the comets of 1531, 1607 and 1682, hypothesizing them to be returns of one and the same comet. Using Newton's laws, he calculated that it would return in 1758, which it did. This was no simple mathematical feat—the gravitational tugs of moving planets frequently alter, by a few years, the *exact* time of return from the *average* time of return.

Hally's most important contribution, however, was recognizing the importance of Newton's insights, and convincing him to write *Principia*. Later, Halley stopped all other work while he concentrated on proofreading and otherwise shepherding the masterpiece into print. There was also a truly major problem—financing its publication. In those days, it was common for a groom to receive a dowry from the bride's family, and Halley used his dowry to pay for all publication expenses.

Teacher. Perhaps not yet, but it turns out that $\dot{\theta}$ can be expressed in terms of r, which will lead to a substantial simplification. Look at Figure 16.7:

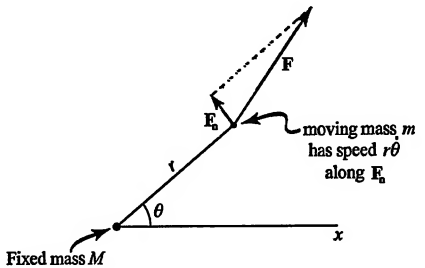


FIGURE 16.7. A force $\mathbf{F}$ creates a torque about M; if $\mathbf{F}_n$ is the component of $\mathbf{F}$ perpendicular to r in the (r, θ)-plane, then this torque is defined to be $r\mathbf{F}_n$. If $\mathbf{F}$ is gravitational attraction, then $\mathbf{F}$ lies in the same direction as r, so $\mathbf{F}_n = 0$. Of course, that means the torque about M is 0.

Torque is the "angular" analogue of force, just as angular momentum is the corresponding analogue of linear momentum, and the two are computed in similar ways. When zero (ordinary) force acts on a system, total (linear) momentum is unchanged. Analogously, when zero torque acts on a system, the total angular momentum is unchanged. Since the speed of m in the direction normal to r is $r\dot{\theta}$, the angular momentum of m is $mr^2\dot{\theta}$. Equivalently, the angular momentum *per unit mass* m is $r^2\dot{\theta}$; if we denote this angular momentum by h, we can write $h = r^2\theta$. As with M, the torque on m is zero, so the total angular momentum of the system is conserved through time—that is, h is a constant. Therefore

$$h = r^2\dot{\theta} \quad \text{implies that} \quad \dot{\theta} = \frac{h}{r^2},$$

which in turn means that (16.7) can be expressed this way:

$$\frac{1}{r^2} = \ddot{r} - \frac{rh^2}{r^4} = \ddot{r} - \frac{h^2}{r^3}. \tag{16.8}$$

Our final step is rewriting (16.8) in terms of u, where $u = \frac{1}{r}$. This of course means expressing $\ddot{r}$ in (16.8) in terms of u and its derivatives. We continue to use the chain rule as needed:

The first differentiation:

$$\dot{r} = \frac{dr}{dt} = \frac{d(\frac{1}{u})}{dt} = -u^{-2}\frac{du}{dt} = -u^{-2}\frac{du}{d\theta}\frac{d\theta}{dt}.$$

Since

$$\frac{d\theta}{dt} = \dot{\theta} = \frac{h}{r^2} = hu^2,$$

this last expression can be written as

$$-u^{-2}\frac{du}{d\theta}hu^2. \tag{16.9}$$

Of course (16.9) simplifies to

$$-h\frac{du}{d\theta},$$

so

$$\dot{r} = -h\frac{du}{d\theta}. \tag{16.10}$$

The second differentiation proceeds in like vein:

$$\ddot{r} = -h\frac{d\left(\frac{du}{d\theta}\right)}{dt} = -h\frac{d\left(\frac{du}{d\theta}\right)}{d\theta}\frac{d\theta}{dt} = -h\frac{d^2u}{(d\theta)^2}hu^2 = -h^2u^2\frac{d^2u}{(d\theta)^2}. \tag{16.11}$$

At this stage we have:

$$u = \frac{1}{r};$$

$$\dot{\theta} = hu^2;$$

$$\ddot{r} = -h^2u^2\frac{d^2u}{(d\theta)^2}.$$

Substituting these into (16.7), which is

$$\frac{1}{r^2} = \ddot{r} - r(\dot{\theta})^2,$$

gives

$$u^2 = -h^2u^2\frac{d^2u}{(d\theta)^2} - h^2u^3. \tag{16.12}$$

Dividing (16.12) through by $(hu)^2$ and rewriting slightly, yields:

$$\frac{d^2u}{(d\theta)^2} + u = -\frac{1}{h^2}. \tag{16.13}$$

Done!

Student. What a *tour de force*! Actually, (16.13) looks sort of familiar—something like

$$y'' + y = K, \tag{16.14}$$

which I know I've seen somewhere.

Philosopher. If prime means differentiation with respect to time, then isn't $y'' + y = 0$ just the famous equation for a vibrating mass-spring setup—a "harmonic oscillator"?

Teacher. That's right, and the constant K just adds an extra constant horizontal force to the setup. The general solution to this is

$$y(t) = \alpha\cos t + \beta\sin t + K.$$

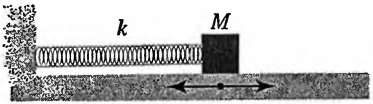


FIGURE 16.8. Newton's great differential equation is close in form to that for the mass-spring: $My'' + ky = 0$. For a long time, Hooke thought that force between two masses varied like the tension of a spring and therefore he spent a lot of time musing about springs. His guess was wrong, but his spring law $F = -ky$ was correct for an ideal spring. The constant k is still called the Hooke constant of the spring.

where α and β are two arbitrary "integration constants" of the linear second-order differential equation (16.14). This solution can also be written in a more geometrically-informative way as

$$y(t) = A\cos(t - \phi) + K, \tag{16.15}$$

where A is amplitude and ϕ is phase shift. In making the transition to (16.15), the two degrees of freedom, α and β, reemerge as the two degrees of freedom, A and ϕ.

A standard exercise in a linear algebra or differential equations course is to prove that the functions $\sin t$ and $\cos t$ are linearly independent over $\mathbb{R}$. It happens that one can in a natural way identify $\sin t$ with the vector $\epsilon_1 = (1, 0)$ and $\cos t$ with the vector $\epsilon_2 = (0, 1)$. Having done this, the vector $a\sin t + b\cos t$ then has rectangular coordinates (a, b). But more than that, it also has *polar* coordinates (r, θ), and the usual connections between rectangular and polar coordinates hold:

$$r = \sqrt{a^2 + b^2} \quad \text{and} \quad \theta = \operatorname{Arctan}\left(\frac{b}{a}\right).$$

This polar form is sometimes called a "phasor" by electrical engineers, and is written $r\angle\theta$. This term is not inappropriate, since $-\theta$ is in fact a phase shift: this is evident from the trigonometric identity

$$a\sin t + b\cos t = \sqrt{a^2 + b^2}\sin(t + \theta).$$

One can establish this by expanding the right-hand side, using the sum-of-angles formula. The right side reveals the geometry lurking in the more algebraic left side, and easily permits one to sketch the graph of

$$a\sin t + b\cos t.$$

Since (16.13) and (16.14) have the same form, so do their solutions, so we can write the solution to (16.13) as $u(t) = A\cos(\theta - \phi) + K$. Since $r = u^{-1}$, we can massage (16.15)

slightly into:

$$r(\theta) = \frac{p}{1 + \epsilon \cos(\theta - \phi)}, \tag{16.16}$$

where

$$p = \frac{1}{K} \quad \text{and} \quad \epsilon = \frac{A}{K}.$$

Philosopher. How very clever! So we have a conic!

Student. Yes, we have a conic, but I thought the whole point was to figure out *which* conic my asteroid is moving on!

Philosopher. He's right. We're supposed to take a reading of the asteroid's position and velocity and get the equation of its path. Then we can sketch the conic and hope that it is comfortably far from earth's orbit.

Teacher. Yes, there indeed *is* that last leg of our journey. Specifically, we should express p, ϵ and ϕ in (16.16) in terms of some simultaneously-observed position and velocity. If these two vectors have polar-coordinates $(\mathbf{R}, \alpha)$ and $(\mathbf{V}, \alpha + \beta)$, then the setup looks like this:

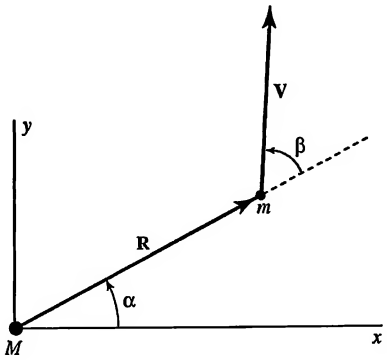


FIGURE 16.9. M is taken to be fixed, and m is the much smaller, moving mass of the asteroid. This picture is meant to record a simultaneous reading of the asteroid's position and velocity.

Working out expressions for p, ϵ and ϕ in terms of these vectors is somewhat involved, but details appear in the literature. (See, for example, [15]). Here are the formulas:

$$p = R^2V^2\sin^2\beta;$$
$$\varepsilon = \sqrt{1 - RV^2(2 - RV^2)\sin^2\beta};$$
$$\phi = \text{tilt} = \alpha - \text{Arctan}\left[\frac{RV^2\sin(2\beta)}{2(RV^2\sin^2(\beta) - 1)}\right]. \tag{16.17}$$

These can be entered into Maple or Mathematica, and once you choose specific values for the position and velocity, either of these programs will easily draw the associated conic in polar-coordinate mode. [Some examples appear in Appendix 1; also, see Appendix Applet 7a on the CD ROM.]

Some Reality Checks

Philosopher. What we've just done is very impressive, but I'm wondering how our results compare with the real world. For example, we assumed that the force of gravity pulling on a planet all came from a single point-mass. But the sun is hardly a point—its diameter is over 865,000 miles! Does that mean that its real-world gravitational effect is different than if the mass were much more concentrated, being closer to a theoretical point? And what about all those other planets in the solar system? As the earth tries to pursue its elliptical orbit, it's got planets like Venus, Mars, Jupiter, Saturn, and Mercury tugging on it! Common sense tells me that none of the solar system planets make exact elliptical orbits. *How far off are our theoretical results from the actual*?

Teacher. Those are salient questions which deserve to be answered.

Your first question actually has a historically important answer, since that very question was raised by Newton himself. Take the earth and sun, for example. According to his law of universal gravitation, every tiny bit of the sun's mass exerts a force on every tiny bit of the earth's mass. If you add up the vast multitude of all these forces, then his question was: would the total force be the same if all of the sun's mass were somehow concentrated at its center? And if that is true, then wouldn't the total force continue to be the same if all the earth's mass were concentrated at *its* center, also? Intuitively, the idea seemed to make sense to Newton. After all, a gram of sun mass closer to us would exert a greater pull on us, and a gram of mass on the other side of the sun, a lesser pull. The combined result might well be equivalent to two grams of mass at the center of the sun. Newton was very cautious and would not make his musings public until this question had been answered. He used his own integral calculus eventually to answer this in the affirmative. The total gravitational pull on a particular planet is the same whether the sun is its actual size, the size of a basketball or the size of a microbe—and likewise for the earth. The only restriction is the commonsense one that we don't alter the center of mass when we change size.

Philosopher. That's reassuring! But I'm afraid that my other concern can't be "proved away" like that. The gravitational influence of all those other planets—each one *must* pull all the other ones away from the theoretically elliptical orbit, right?

Teacher. Yes. But as you suggested, the big question is *how much*? The answer is *very little*. Take Mercury, for example. It turns out that due to its relative lightness as well as its position in the solar system, it is the planet most prone to be whiplashed by some of the

much more massive ones. And in fact it's affected not only by the other planets, but also by general relativistic effects, which Newton knew nothing about. All this means that instead of its path closing up perfectly after an orbit around the sun, it slightly misses it, leading to **precession**. Figure 16.10 shows an exaggerated example of this. In the left picture, a planet has started traveling counterclockwise, suggested by the arrow. After one orbit, the planet's velocity has changed direction significantly, and the picture shows it nearly completing a second orbit. The picture on the right shows a continuation of this process.

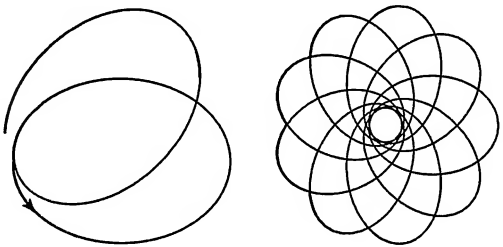


FIGURE 16.10. In this exaggerated view of planetary precession, the major axis is rotating 40° per orbit. In general, the path won't close up and retrace itself, as it does here.

The actual rate of Mercury's precession is actually less than *one hundred thousandth* of that drawn in Figure 16.10, the major axis rotating a mere 0.000384° per orbit!

> Suppose we make a computer simulation of Mercury's orbit that takes into account the influences of all the other planets as well as relativistic effects. If we speed up the simulation so Mercury completes one orbit not every 88 days (its actual rate), but rather one orbit each second, then even after running an hour, the amount of turning would be less than one degree. To precess a full 360°, the simulation would have to run continuously for over $11\frac{1}{2}$ days, representing an actual time of more than 225,000 years. And, as it turns out, Mercury experiences the fastest deviation of all the nine planets. The behavior of the others is even better.

Student. I think I have found something wrong with this concentrated-mass business. You're saying that if a planet, like the earth, suddenly shrunk to the size of a baseball, without losing any mass, then orbits wouldn't change?

Teacher. Certainly!

Mercury's precession figured prominently in the work of both Newton and Einstein. Newton proposed that observing Mercury's precession rate would serve as a sensitive test of his universal gravitation law. And it did: his law accounts for 531 seconds of turning per century, leaving only 43 of the observed 574 seconds unaccounted for. But that's still 43 seconds unaccounted for! *What could be the source of the discrepancy*? The mystery stubbornly persisted for over two centuries, yielding only to Einstein's general theory. Those 43 seconds come from general relativistic effects, Mercury's precession once again offering a crucial test of a universal theory.

Student. OK, suppose I throw a stone. The stone follows a parabolic path— Galileo proved it, and I got a 100% on an exam where I had to prove it, too!

Teacher (Getting a little nervous about this).

Student. Now, when my stone is flying in the air, I suddenly shrink the earth to the size of a baseball. If the stone's path due to gravity isn't affected, then the stone doesn't know anything about the earth suddenly shrinking, and the stone continues on a parabolic path, not in some elliptical orbit. Now a parabolic path means my stone is going to escape from the solar system, when in actuality hundreds of thousands of similar stones are trapped, circling planets, forming rings around Saturn and other planets, and so on. You're saying that shrinking the earth does nothing at all to any of these other planets and rocks, yet mine is free! Something seems wrong.

Philosopher. *Wow*—he certainly has a point! After all, he didn't expend very much energy throwing it. You'd hardly expect that he managed to catapult a stone so that it ultimately escapes the earth's powerful and relentless gravity—a one-man launching pad!

Teacher. (Trying to think quickly, here.) Well, well, that is indeed a question... what I'm thinking is that the earth is normally so huge that the stone crashes into it instead of going into orbit. I think that if the earth actually did suddenly shrink to baseball size, then the stone probably wouldn't crash into the very-dense baseball, but would in fact travel around in an orbit.

Student. You're saying I *would* be a one-man launching pad?

Teacher. Yes, but that's because of one tremendous advantage: when you launched it, you were 4000 miles from the center of the earth—that is, 4000 miles from the center of the baseball. To launch it there, you in effect made a 4000-mile climb up a ladder, from the baseball-sized earth to your launching position, and you did a lot of work to get there. From a celestial laboratory that far up, you could indeed make all sorts of launches. Your stone, when it's near the surface of the baseball-sized earth, would be *incredibly* heavy since the denominator r in Newton's inverse-square law is so small. Of course way up there—on the surface of actual earth—it weighs the usual, ordinary amount, since r is then much larger.

Now that orbit around a baseball with the earth's mass would periodically lie outside an imaginary earth-sized sphere for just a few seconds. All the rest of the time the orbit would lie within the sphere.

Student. But now you're talking about elliptical orbits! The path is actually a parabola.

Teacher. I have to disagree with that.

Student. *Disagree*?! But my exam—my 100%. And Galileo...

Teacher. I think there's a way to make peace on all sides here. Galileo assumed that the force due to gravity is constant, and that is justified for the relatively short heights and distances traveled by a stone or other everyday-sized projectiles on our normally-sized earth. Arguing under that assumption, Galileo did correctly show that the path must be precisely a parabola. But over even those small distances, the gravitational force does in fact vary a tiny bit, and if we shrink the earth so it doesn't rudely intercept the stone and allow the stone to continue on unimpeded, then after a time this variation becomes greater, and the stone increasingly feels that variation in gravitational force, which of course depends on its distance from the shrunken earth. When you toss a stone, you actually see just a tiny bit of an *enormous* orbit, near a major vertex. Because of the nearly constant force, that little part is a close approximation to Galileo's parabola.

Student. You mean we and Galileo are *both* right?

Teacher. Yes, both answers are right, both being logical consequences of the assumptions that were made. *We* are using one set of assumptions, taking a larger-scale view. *Galileo* used a different set of assumptions, viewing the world on a more everyday scale. The inverse-square law hadn't been discovered in Galileo's time, so his assumption was not only very nearly correct, but also a very understandable one.

Let's assume that we've tossed the stone, and that the earth is then magically compressed to the size of a baseball. What can we say about the stone's journey? First, the stone's elliptical orbit is extremely elongated. After rising 20 or 30 feet, the stone has gone as high as it's going to, and then it begins to come down again. Without the normal earth in the way to stop it, it continues to accelerate for 4000 miles toward the center of mass at the center of the shrunken earth, and because the earth is so small, the stone misses it. But it's traveling at a blinding speed.

Student. So it whizzes past the dense baseball and continues on "rising" an equal distance beyond it?

Teacher. No—it goes just a small distance beyond the dense baseball.

Student. *What*?! I can't believe that. It seems obvious that the great speed would make the stone continue on for 4000 miles, turn around, and keep doing that.

Teacher. Remember, the earth's center is at a *focus* of the orbit, and our baseball-sized earth is so small that the stone doesn't crash into it!

Student. Somehow, I just can't feel that's really true.

Teacher. Well, look at it this way: when the stone is extremely close to the dense baseball, there's a *fantastically large force between them*, because gravitational attraction is inversely proportional to the distance-squared between baseball and stone. The denominator is small and therefore the force, huge. As the stone tries to flee the baseball, this tremendous force whips the not-very-massive stone around almost 180°, forming an arc that looks exactly like the graceful arc your stone made when you threw it, though now it's moving *tremendously* faster.

Student. The force is really *that* strong?

Teacher. We can do the math. We're talking about 4000 miles versus, say, maybe just a foot, which is a ratio of $4000 \cdot 5280 = 21{,}120{,}000$ to 1. But it is the ratio of distances *squared* that enters into the inverse-square gravitational law—and that gives us a nearly 4.5 *trillion* (million million) to 1 ratio! An ordinary stone weighing one ounce is experiencing a force of nearly 14 billion tons.

Student. I guess you're starting to convince me.

Teacher. The baseball is acting like a star, and the stone is playing the role of a highly-elliptical planet. One focus of the ellipse is at the center of the baseball, and the major axis of the elliptical orbit is essentially the radius of the normal earth—about 4000 miles. That means its semi-major axis is 2000 miles long.

Philosopher. What you just said doesn't prove that the ellipse is extremely elongated, does it? Just big.

Teacher. But I can estimate the semi-minor axis, too, getting the shape.

Student. How?

Teacher. Consider the highest point reached by the stone. What size circle do you think best fits its path there?

Student. Oh, maybe a circle a yard or so in diameter.

Teacher. Sounds reasonable. The curvature at that vertex is the reciprocal of your circle's radius of 1.5 feet. We also have a formula for the curvature at this vertex—it's

$$\kappa = \frac{a}{b^2}.$$

Student. Where does that lead us?

Teacher. κ is measured in $(\text{miles})^{-1}$. Now 1.5 ft converts to $\frac{1}{3520}$ miles, so $\kappa = 3520$ $(\text{miles})^{-1}$ We know $a = 2000$ miles; substituting into the formula

$$\kappa = \frac{a}{b^2}$$

gives

$$3520 = \frac{2000}{b^2}, \quad \text{or} \quad b = \sqrt{\frac{2000}{3520}}$$

miles. So b is roughly a half mile. The ellipse is about 4000 miles long and a mile wide.

Student. That really *is* elongated, isn't it?

Teacher. It's a ratio of about 1 to 4000. A sketch 40 inches in length would lie in a rectangle only $\frac{1}{100}$ inch wide. Our eyes would simply see a straight line segment. The eccentricity of this ellipse is

$$\frac{\sqrt{2000^2 - \frac{1}{2}^2}}{2000} \approx 0.999999968.$$

Sketching the polar equation (16.2) [p. 315] with $\varepsilon = 0.999999968$ versus that equation with $\epsilon = 1$, our eyes—as well as even quite accurate measuring instruments—couldn't tell any difference between them for quite a distance around the origin.

> When you throw a stone, or gaze upon the graceful arch from a water fountain, you're witnessing a truly remarkable sight: a close-up view of part of a *tremendously* long and relatively thin ellipse having "radii"—or principal semi-axis dimensions—of approximately 1 mile and 4000 miles. We're looking at a relatively small neighborhood of a vertex, within which the ellipse looks very much like a parabola. Because we can't visually tell any difference between our limited view of a parabola versus that of this very thin, monstrous ellipse, we're justified in using the misnomer "parabola" to describe the path.

Student. You said the stone would whip around the dense baseball in a path looking exactly like the arc it started off in, but much closer, and much faster. I'm curious—*how close*, and *how fast*?

Teacher. The answers are actually pretty easy to find. We found that for the ellipse, $a = 2000$ miles and $b = \frac{1}{2}$ mile, so the nearest distance is $a - \sqrt{a^2 - b^2}$ miles. That works out to an incredibly tiny *four inches*! It barely misses the baseball.

Student. Wow! We'd better assume our stone is pretty small, like an inch in diameter.

Teacher. A good idea. That would leave two inches between the spherical stone and a three-inch diameter baseball. This is because the stone has radius 0.5 inches, and the baseball has radius 1.5 inches. When they are at their closest, their surfaces are two inches apart, since $2 = 4 - 0.5 - 1.5$.

Philosopher. But wouldn't the stone's center come even *closer* than four inches? We just saw that the ellipse looks like a line, so just a little way from the vertex, the stone would get even closer to the focus, right?

Teacher. No, it wouldn't.

Philosopher. Why not?

Teacher. Well,

$$r = \frac{p}{1 + \varepsilon \cdot \cos\theta}$$

is a polar equation of an ellipse with the pole at the right focus, and the form of this fraction makes it clear that r assumes its minimum at $\theta = 0$.

Philosopher. In any case, I doubt the stone would survive such a trip without breaking up into tiny pieces...

Student. *Whew*—I think I've learned a lot! By the way, I get to keep my 100%, right?

Teacher. It's OK with me!

Another Example of Inverse-Square Force

Philosopher. If Newton's laws are actually universal, then couldn't a 500-pound asteroid floating around in the universe have a few pebble-sized "moons" orbiting it? Maybe even *rings*, like Saturn?!

Teacher. A small chunk of uncompressed mass is not likely to have anything orbiting it.

Philosopher. Why not?

Teacher. Gravity is an extraordinarily weak force; the force generated by a small chunk is so weak that it would take very little speed to escape from a theoretical orbit around it.

Student. What do you mean, *weak*?! It's really hard to launch things into orbit because gravity is so strong. It takes so much energy, that the rocket engines literally shake the ground. I've heard that people who've actually seen and felt this are almost always unprepared for the experience and never forget it. Besides, anybody falling off a roof certainly feels gravity! It seems to me that gravity is awesomely strong.

Teacher. The gravity we experience in daily life is the combined gravitational force exerted by *all the rest of the earth*. But a little chunk of asteroid creates hardly any gravitational tug at all. Interestingly, however, if you really *compress* it—keeping the mass fixed—then the inverse-square law gives us the potential to create large forces at small distances. We just saw that with the stone/baseball example. If you're able to make the denominator small in $\frac{GMm}{r^2}$ without changing the numerator, you can get a large attractive force, large enough to hold a point-mass satellite in orbit. An example is electric charge, which also obeys an inverse-square law. And, as it turns out, electric forces are incredibly stronger than gravity. You again get orbiting, though at a much faster rate.

> Whenever the force between particles obeys an inverse-squared law—such as gravitational or electrical force—the orbiting particles obey Kepler's Laws. But their orbiting rates encompass an astonishing range. For example, our solar system orbits the Milky Way Galaxy's central black hole (which has a mass of over 2 million suns), taking about 230 million years to complete one trip. The earth circles the sun once a year; our moon orbits us approximately once a month. In the classical orbital picture of hydrogen, the electron goes around the central proton about 1 quadrillion (10^{15}) times every second.

Student. Incredibly stronger than gravity? I know there are electrons and protons in every atom of matter, and that means you've got the whole earth load of *them*, yet it's only during a dry winter day that I'm occasionally aware of a little bit of static electricity.

Teacher. Yes, but remember that electricity comes in two flavors—positive and negative, and the forces arising from these charges almost always perfectly cancel out. If they don't, then we can see daunting power.

We've all witnessed the tremendous energy created when even a tiny weight of positive charge gets separated from negative charge: lightning. During a thunderstorm, violent air currents slam ice crystals or water droplets together, doing so with such force that trillions of electrons get ripped off from their atoms. The churning air serves as a kind of natural atom smasher. These torn-off electrons tend to gather at the bottom of the thundercloud, and a corresponding region of positive charge gets created in the ground below. This represents a very minute separation of positive and negative charge—so small that the excess electrons total up to less than *one billionth the weight of a penny*. Nevertheless, a voltage of several billion volts builds up, creating a "fatal attraction" between cloud and ground charges. When it comes, the stroke of lightning heats up the surrounding air, forming a plasma having a temperature some five times that of the sun's surface. The super-hot particles scream outward at supersonic speeds, creating a large sonic boom which we hear as thunder. The amount of energy released in one thirtieth of a second can exceed an entire year's electricity consumption of a small city.

As an example of the tremendous difference in strength between gravitational and electrical forces, think of all those little dust particles that gather on a TV screen. There's normally just a minute excess negative charge on the screen, a byproduct of the cathode tube's shooting electrons at the screen phosphors to create a picture. But that minute excess negative charge makes those dust particles cling to the screen. And what is so remarkable about that? The combined gravitational force of an *entire earthful of mass, all of it tugging together to pull the particles downward*, can't budge them. Almost all those bits of dust would cling there indefinitely if you didn't wipe them off yourself!

Philosopher. So there's a whole other world out there, one that most people are unaware of. Are you suggesting that what we've learned about gravitational orbits also applies to electric charge?

Teacher. To an extent. Because electrical forces are so much stronger, the distance scales are very tiny. Though the "electron orbiting a proton" model of the hydrogen atom is more appropriately described in probabilistic terms, in this planetary picture the electron orbits some 10^{15} times a second, and that is, not coincidentally, about the frequency of visible light.

Center vs. Focus—Again

Student. To be honest, I still have a problem with where the sun is. I know the sun is at a focus of each elliptical planetary orbit, but I still just don't feel right about it. To me, it really seems that the sun should be at the center of all these orbits!

Philosopher. I have to confess, that *does* seem more symmetric.

Teacher. But consider this: suppose the sun were gradually to lose mass, and a planet were to gain mass. Then the original predominance of the sun would begin to diminish.

This could continue to the point where the sun and planet have equal masses! *Then which one is king*? Neither one is dominant, and it would be unfair to say that one should be the center of the other's orbit. In fact, using a focus represents not only symmetry, but a deeper and more "democratic" symmetry than the center.

Philosopher. Can you express this symmetry in a more quantitative way?

Teacher. Well first, there *is* a "king," but it's not the sun, but rather a *location*: the system's *center of mass*—the center of mass of the two bodies. It is *that* point which remains fixed, and each body orbits around it. That fixed center of mass is a focal point for each orbit.

Student. But the planets orbit around the sun. You're not saying the *sun* orbits, are you?

Teacher. Yes, I am. Because the sun is so much more massive than the earth (the sun accounts for over 99.8% of the entire solar system's mass), the fixed focal point is close to the sun's center—so close that the fixed point actually lies *deep within the sun's fiery ball*. Suppose we consider just the sun and the earth, and let's shrink the sun to the size of a basketball. The earth orbits it in a near-circle of radius about 93 million miles. That enormously heavy basketball would travel around the fixed sun-earth center of mass in a nearly circular orbit of radius about 280 miles, or diameter 560 miles. When you return the sun to its actual size, its *center will continue tracing out exactly that same orbit* of diameter only 560 miles. Now the sun's actual diameter is about 870,000 miles. Suppose we make a model in which the sun is a sphere one foot in diameter. Then what would be the diameter of the 560-mile circle? Answer: $\frac{560}{870,000}$ feet, which is less than a hundredth of an inch! So the moving sun would look like a soccer ball one foot in diameter, with its center moving around that very tiny circle.

In contrast, suppose you look at a binary star. If one star is twice as massive as the other, then the center of mass is at the one-third, two-thirds point between them. Calculating the center of mass is done the same way as finding the balance point of a seesaw—you equate moments. So if a 150-lb. adult balances a 75-lb. child, and if the child's seat is 12 feet from the fulcrum, then the adult is 6 feet from the fulcrum: $75 \times 12 = 150 \times 6$.

Philosopher. So there are *two* different elliptical paths!

Teacher. Yes, but things are very beautifully symmetric. We see naturalness in the shapes of the two ellipses, in their size, mutual tilt, location—there's a natural symmetry throughout. If the bodies have masses M and m, then here are the facts:

- Shape: both ellipses have the same shape—that is, the same eccentricity. In the terminology of Euclidean geometry, the ellipses are *similar*.
- Size: ellipse sizes behave like fulcrum distances in that

$$M \times (\text{Diameter of } m\text{'s orbit}) = m \times (\text{Diameter of } M\text{'s orbit}).$$

- Mutual tilt: their major axes lie on the same line.
- Location: the fixed point is the right focus of one ellipse, and the left focus of the other ellipse.
- Both masses orbit with the same angular speed and therefore have the same period.

This picture incorporates a lot of what we've just said:

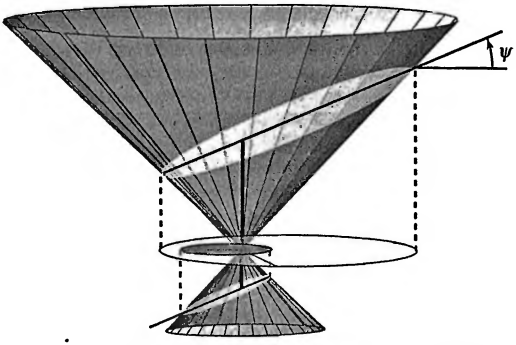


FIGURE 16.11. Two parallel planes intersect a circular cone in similar ellipses. The angle ψ is determined by the choice of initial conditions—how the body was "launched" into orbit. For example, choosing $R = V = 1$ and $\beta = 90°$ in (16.17) on p. 326 yields $\psi = 0$, giving two concentric circles. The two parallel planes intersect opposite nappes of the cone; the cone's vertex is the right focus of one projected ellipse, and the left focus of the other one. The "altitude" of a plane's intersection with the z-axis is proportional to $\frac{1}{Mass}$. The sun thus has a smaller altitude, so its orbital plane cuts off a smaller ellipse. A light planet has a greater altitude, so its orbital plane cuts out a larger ellipse.

Here's another look at the two similar elliptical orbits:

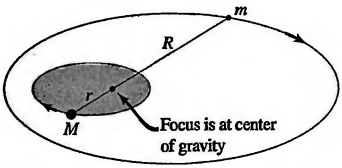


FIGURE 16.12. In the two similar ellipses, the line segment from M to m rotates around the fixed center of mass which, in this picture, is the right focus of the smaller ellipse and the left focus of the larger one. When the line through M and m line is horizontal, then either M or m are simultaneously closest to the center of gravity, or they're simultaneously farthest away from the center of gravity. In fact, M and m constantly "pantomime" each other, M uniformly scaling up to "draw" m's orbit, like an enormous pantograph.

Here are some orbit shapes, for increasing eccentricity versus increasing ratio $\frac{m}{M}$:

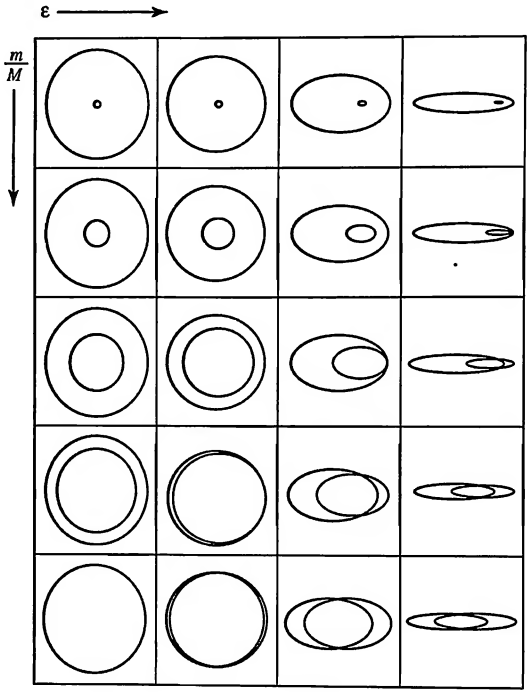


FIGURE 16.13. This is a matrix of orbit shapes for the two-body problem. As the eccentricity increases from 0 toward 1, the two similar elliptical paths become ever more elongated. As the ratio $\frac{m}{M}$ of masses increases from nearly 0 to 1, the similar ellipses approach congruence. In each window of the bottom row, the two ellipses are congruent—in fact, they're translates. In every matrix window, the right focus of one ellipse is the left focus of the other.

Philosopher. If only Kepler had known! Today we have a better sense of symmetry. Just suppose Kepler were actually in the room with us today. He'd given up on circles and had to live with the imperfection of ellipses. It seems to me that all these pictures would make him feel better.

Chapter Summary

Historians believe that the earliest applications of conics by the ancients were to obtain approximate solutions to "unsolvable" problems, such as squaring the circle, doubling the cube or trisecting an arbitrary angle—in fact it's surmised that conics were first considered for this purpose. The chapter begins by illustrating this for the cube-doubling problem, showing how the ancients approximated a parabola. (Of the conics, only the circle and degenerate ones are exactly Euclidean-constructible.) Chapter Applet 23 on the CD-ROM shows an animated version of the construction in Figure 16.3.

During the Dark Ages, conics faded into obscurity. In attempting to determine what paths planets travel in, Kepler eventually realized that Mercury's path had to be some sort of oval. At one point he chanced upon that almost-forgotten curiosity, the ellipse, as the curve that so closely matched Tycho's data for Mercury's orbit. Through the years, Kepler's discoveries became distilled into his "Three Laws of Planetary Motion," which was a quantum leap for the "celestial physics" of his day. Around that time, Galileo studied "terrestrial physics," and proved that the path of a cannonball or other projectile is a parabola, contradicting some of Aristotle's teachings. These discoveries pulled conics out of their dark-age obscurity and, in the hands of Newton, led to the most famous-ever application of conics: classical astrophysics.

It was Newton who unified celestial and terrestrial physics into one grand, universal theory. Kepler's laws, it turned out, represented only part of the story—what might be called the "low energy case," where a planet doesn't move fast enough to escape orbiting the sun. With enough speed, however, any heavenly body will escape and drift endlessly away along a parabola or one branch of a hyperbola. Newton approached the problem as we often do today, describing local behavior through differential equations and then solving them to arrive at the entire range of possibilities. The chapter derives in this way the motion of a (point) mass moving under the influence of the gravitational field of a fixed mass. Newton proved that in his physics, *the path of a point mass in a centrally-symmetric gravitational field is a conic.* Conics now played a central role in Newton's great scheme of the Universe.

His great applications of conics to the Universe assumed *point*-masses. But the earth and sun are hardly points! Importantly, Newton used his own discovery of the calculus to prove that the mass in a large ball like the sun, earth or moon can be theoretically concentrated to a smaller ball or an ideal point without affecting the gravitational influence on objects outside the ball. In this chapter, we use this fact to link terrestrial with celestial physics: in a "thought experiment," we throw a stone and then suddenly shrink the earth to the size of a baseball. The stone never realizes what has happened, and finds itself traveling in an elliptical orbit about the shrunken earth—an image as thoroughly celestial as it is terrestrial. Looked at this way, our moon is just a large stone. What about Galileo's parabolic projectile path? That is correct only if one assumes a constant gravitational field, which is a very good approximation in the special case of everyday life on the earth's surface. Over larger distances, however, variation in gravitational strength becomes so significant that in our example, what initially looks like a parabolic arc turns out to be an elliptical arc.

We end the chapter by clarifying Kepler's first law, which puts the sun's center at an ellipse focus. Once again, a very good approximation, but the deeper and more symmetric

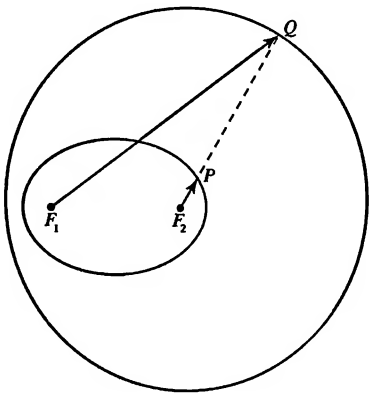
truth is that the earth and sun each orbit around their center of mass, which is the common focus of two similar ellipses.

Exercises

1. At one point, Philosopher seems to be musing that if the earth were compressed to the size of a baseball, the stone would never survive a round-trip in one piece. What do you suppose he is thinking?
2. Kepler's second law is really just another way of saying that the vector quantity of angular momentum is conserved.
   (a) Show this.
   (b) Does this mean that an asteroid sweeping past the solar system in a hyperbolic path also obeys Kepler's second law?
   (c) Suppose that gravity were somehow suddenly switched off. In what paths would the planets then move? Would they continue to obey Kepler's second law?
3. (a) Venus has the most nearly circular orbit of all the sun's planets: its eccentricity is only 0.0068. Its average distance from the sun is about 72.3% of our earth's average distance. Use Kepler's third law to calculate how many earth-days in one Venusian year.
   (b) Neptune's orbit eccentricity is 0.0086, next to Venus in "circularity." Neptune is approximately thirty times more distant from the sun than we are. How many earth years are there in a Neptunian year?
4. If a planet moves in an ellipse of eccentricity ε, what is the ratio of the maximum to minimum speeds?
5. Suppose a small but potentially very destructive asteroid is heading directly toward us, and that we have launched an explosive device to knock the asteroid out of its collision course. The device is timed to detonate as it passes nearest the asteroid, thus giving the asteroid velocity perpendicular to its path. If the asteroid is one million miles from us and is traveling at 1000 mph, what minimum speed must the device impart to change its orbit to a hyperbola, so that it will never return to haunt us?
6. Hooke once thought that the force of gravity increases directly with r, as if each planet were attached to the sun with an invisible spring obeying, of course, Hooke's linear spring law, $F = -kx$. (Still today, a spring's strength (or "stiffness") k is often called its "Hooke's constant.")
   (a) For a planet moving in the (x, y)-plane, write one differential equation for the x-coordinate of its motion, and another one for the y-coordinate. (These are *far* simpler than Newton's correct ones!)
   (b) Show that the motion is indeed elliptical.
   (c) Show that the sun would be at the center of the ellipse instead of at a focus.
   (d) Does Hooke's theory rule out hyperbolic motion?
7. Suppose we could take a stone weighing 1 ounce here on ordinary earth, and actually walk it up from the surface of a baseball-sized earth to where you're reading this book. How much work (say, in ft-lbs) would that take?

8. Make a rough estimate of the eccentricity of the ellipse created by suddenly shrinking the earth to baseball size for:
   (a) a golf ball rising 100 feet in a 300-yard shot.
   (b) the arc of water in an ordinary drinking fountain.

9. From Kepler's first law, we know that the sun is at one focus of the elliptical orbit. But *Philosopher*'s symmetry-seeking radar is starting to beep—that law singles out *one* focus. He's beginning to wonder, *Is there any significance to the other one?* It seems to him that Nature would never just leave a "free focus" dangling, having no importance. There is an answer, and in fancy language, it's "conjugate foci determine conjugate dynamical quantities." The conjugate dynamical quantities are just position and velocity. The answer can also be given pictorially:



Here, the sun is at focus F_2, with planet P orbiting it. The circle is centered at F_2, and its radius is the ellipse string length—that is, its major axis. (One can switch from ellipse construction to circle construction by pulling out the stake at F_1 and using the string anchored at F_2 to trace out the circle.) The vector from F_2 to P gives P's *position*, and in appropriate units, the vector from F_1 to Q "almost" gives P's *velocity*—one need only rotate the vector 90°. One may philosophically ask why one must rotate by 90°. P *does* specify the planet's location, but it doesn't say whether it's orbiting clockwise or counterclockwise. Nature seems to be doing what it can to maintain neutrality here, and once we know the sense of rotation, it's clear which way to rotate the velocity vector. The above picture for the sun-planet system is actually the union of two diagrams: the ellipse and position vector F_2P form what physicists sometimes call the system's "position diagram"; the circle and velocity vector F_1Q form the system's "velocity diagram."

(a) Find a formula in terms of t for P's speed (which we assume to be the length of the velocity vector) when P is at $x = a\cos t$, $y = b\sin t$, where $a > b$.

(b) P's velocity vector is tangent to the ellipse at P. Prove that this vector is in fact perpendicular to the vector from F_1 to Q.

(c) Assuming Kepler's Laws, justify that P's velocity vectors based at F_1 really do describe a circle. Which one(s) of Kepler's Laws are actually used to make this justification?

10. Let E be the elliptical path of a planet orbiting a fixed sun, and let the planet's period (or "year") be P. Starting at any point on E, mark three points on E at times 0, $\frac{1}{3}P$, $\frac{2}{3}P$ The three tangents to E at these points define a triangle with vertices a, b, and c. Place this diagram anywhere you wish in the complex plane $\mathbb{C}$; a, b, c can then be looked at as complex numbers, which are in turn the zeros of the polynomial $q(x) = (z-a)(z-b)(z-c)$. *Show that the ellipse foci are the zeros of the derivative* $\frac{dq}{dz}$, *and that the ellipse center is the zero of the second derivative* $\frac{d^2q}{dz^2}$. (Other interesting facts: the sun is at the centroid of the triangle; this centroid and E together trisect each of the three triangle medians; and E is the largest-possible ellipse lying within the triangle. And, finally, a bit of perspective: $\mathbb{C}^2 = \mathbb{R}^4$ has four pure "species" of plane—Euclidean, anti-Euclidean, Minkowski, and complex. In this book, we've seen how the familiar real ellipse shape naturally appears in each of the first three. Now we know it also arises naturally in $\mathbb{C}$!)

17

Waves and Conics

in which interfering waves generate hyperbolas, making very precise measurements possible

Waves Make Hyperbolas

Student. I've noticed one big difference between ellipses and hyperbolas—it comes up when you use the stakes and string method. It is easy and fun to draw an ellipse: use a pencil to make the string taut, and then just swing the pencil around to trace out a nice ellipse. Actually, I saw this done for the first time when I was a boy—someone building a house had driven two stakes into the ground and fastened a long piece of twine to them. He used a can of spray paint to keep the line taut, and slowly walked along, creating a big orange oval to mark out a "circular driveway," which was actually elliptical.

Although holding constant the *difference* in string length is fine mathematically, there doesn't seem to be any natural, easy physical way to do this to make hyperbolas.

Teacher. Nature seems to have hit upon a creative solution to your question. Suppose you throw two stones into a perfectly quiet lake. Each stone creates a circular ripple whose radius increases at a constant rate. The stones usually don't hit the water at exactly the same time, and if you say that each circle is "born" the instant the stone hits the water, then one circle is "older," the other younger.

Now these two circles continue to grow, or "age," but their age difference of course always stays the same. After a time, they meet—first at a double intersection point, and then in two symmetrically-separating points. Basic physics tells us that the signed heights of ripples add when they meet, and that means that the two separating points are seen as crest-points of greater height. As these "high-altitude points" move, they trace out a hyperbolic locus.

Animated Applet

For an interactive version of Figure 17.1, next, see Chapter Applet 24 on the CD-ROM.

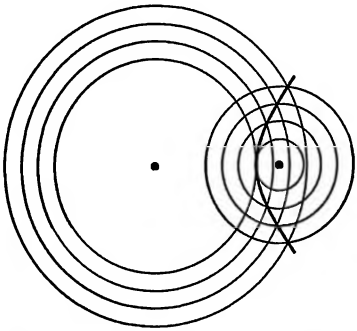


FIGURE 17.1. These circles represent a multiple-exposure photograph of two circular ripples created when two stones are tossed into the water at slightly different times. The stones hit the water at the two central dots, the left one being hit first. The earliest of the four exposures appears as the smallest circle about each dot; it was taken as the two ripples first touch. Their intersection is a double point at the vertex of a right hyperbola branch. We've drawn in this branch on the photo. In succeeding exposures, we see that the double point has separated into two single ones lying on that branch. All this works because the two circles expand at the same rate.

Student. I admit that *looks* like a branch of a hyperbola. Did you just draw it to look that way?

Teacher. That really *is* a hyperbola branch!

Student. How can you be sure?

Teacher. Take the foci to be the points where the stones hit the water. Because the two circles expand at the same rate, we can use their radii to measure their "age." Therefore, if the radius (the "age") of the first circle was one unit when the second one was born, then their radii continue to differ by 1. Of course, each radius is just the distance from the two points of impact—that is, from the two foci. The expanding circles automatically maintain a string-length difference of 1, making the locus a hyperbola!

Of course, that constant difference can be anything you want, depending on when the stones hit the water.

Philosopher. It seems, by symmetry, that if the first stone were dropped after the second one, you'd get a *left* hyperbola branch.

Student. I have to say, nature is pretty clever!

Teacher. And deep, too—nature actually goes a step further than our experiment. Let me explain. Instead of simply dropping two stones, suppose at each location one continuously creates ripples. This is not at all an artificial idea—a star, for example, continuously radiates light waves. To make our water-wave example simple and concrete, let's suppose each stone steadily oscillates at ten full cycles a second; and to keep everything simple, let's also suppose they oscillate with the same intensity, or amplitude. These two stone oscillators

therefore produce two *wave trains*. Since any two expanding circles determine a hyperbola branch, the two trains, as they intersect, create a family of confocal hyperbolas.

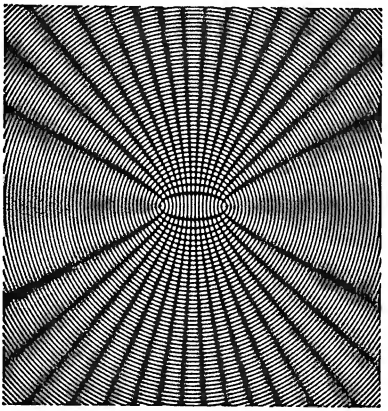


FIGURE 17.2. Here, both stones oscillate at the same frequency, have the same amplitude and are in phase. The two wave trains intersect in a confocal family of standing-wave conics. Light and dark hyperbola-shapes are quite evident; we also see, less distinctly, confocal ellipses. The water is perfectly calm along hyperbola-shaped nodes lying along the corridors between successive light hyperbola branches.

Student. I saw this once in a physics class demo. The teacher used an electric tuning fork with a needle soldered to the end of each prong. When the vibrating needles touched the surface of a vat of still water, they created two sets of circular ripples which you could hardly see because the needles were vibrating so fast. But the fringes were very clear.

Philosopher. That's a fantastic example!

Student. To me, however, those fringes looked straight. I never realized they had anything to do with hyperbola branches.

Teacher. Let's look at some particular point P in any one of the hyperbolas where two crests meet at time t. Assuming there are ten full oscillations per second, then at time $t+\frac{1}{20}$ we see at P not two intersecting crests, but two intersecting *troughs*. Because water-height displacements add algebraically, the trough there is twice as deep as either circle's trough, in the same way that the intersecting crests were twice as high. So there's a lot of activity on these standing-wave hyperbola branches. But in between...

Student. Life is different there?

Teacher. Yes. Look at where a crest from one radiating source meets a trough from the other source. Since displacements add algebraically, we now get *zero* change in water

height. So on *those* hyperbolas, life is serene—for example, a mosquito could skim along such a curve with no problem.

Student. Sort of like living in the eye of a never-ending hurricane!

Teacher. Right. These paths are one-dimensional analogues of the zero-dimensional "nodes" of a vibrating string, which can be demonstrated nicely on a violin or cello, for example. Carefully draw the bow across a string, but don't otherwise touch the string. Now, as you continue to draw the bow, lightly touch the vibrating string at its midpoint, and the note will be an octave higher. If you then remove your finger from the string and continue to draw the bow, the higher octave will continue to sound! You've created a "standing wave," and the nonvibrating midpoint is called a node. You could lightly touch the string at either its one-third or two-thirds point to create a series of three standing waves with two nodes; or at the quarter or three-quarter point to make four standing waves with three nodes; and so on.

In Figure 17.2, the one-dimensional nodes are the hyperbolas in the centers of dark fringes, along which the water surface is perfectly quiet. The other hyperbolas run along standing waves of varying amplitude. If you look along the center of a light fringe, you'll see a hyperbola-shaped standing wave. It's vibrating like a long, curved violin string sounding at a high pitch, having closely-spaced nodes.

From the standpoint of light waves, the calm areas are regions of zero net electromagnetic activity. No light emanates from those areas, so they appear as dark. Maximum oscillation areas, on the other hand, are bright. These "interference fringes" form the basis of some of the most astonishingly precise measurements ever performed.

Student. This is starting to sound exciting! What measurements are you talking about?

Teacher. Let's bring a bit of perspective to things. Normal human eyesight can resolve an angle of about one minute—one sixtieth of a degree. That is, we can resolve (see as distinct) two points of light subtending an angle of one minute at our eyes. In terms of a physical experiment, someone with 20/20 vision, from three miles away, can look at two 100-watt light bulbs five feet apart, shining in the darkness, and see that they are separate lights. If they were brought closer together, they'd blur—someone with a "mere" 20/20 vision couldn't be sure whether the light source came from one point of light or two.

> Around 1950, after several centuries of telescope making, our best telescopes had achieved the ability to resolve an angle of one second, which is one sixtieth of a minute. That's like discerning from three miles that two points of light are distinct when they are separated by a mere one inch. The Hubble Space telescope can see those same points as distinct even if you put them only $\frac{1}{8}$ inch apart. Using the phenomenon of interference fringes—interferometry—we could put those points a *hundredth* of an inch apart, and still see them as distinct from three miles away. By 2010, we expect the resolution to be so high that one can distinguish individual hairs on an astronaut standing on the moon!

To see concretely how one uses oscillation at focal points to obtain such magnification, let me tell you about another piece of magic in the wave story. It was published around 1690 by the physicist Christiaan Huygens (1629–1695), and his principle says that *each point of an advancing wave front itself acts like a miniature wave source.* Figure 17.3 gives the basic idea.

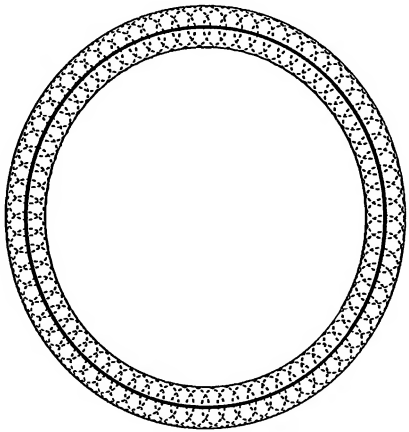


FIGURE 17.3. This illustrates Huygens' Principle, which allows one to construct the shape of a moving wave front at a slightly later (or earlier) moment. The heavy circle between the other two large ones represents an expanding wave front. The Huygens' Principle says that we can consider each point of that front to be stationary, each such stationary point acting as a tiny wave-source. One can imagine a circle of pebbles oscillating in unison. All those small dotted circles represent child-waves recently generated by each pebble, and each of their radii increases at a uniform rate. Now the *envelope* of a continuous family of waves is the curve tangent to each wave of the family. In this case, the envelope of the family of child-waves consists of the two other large circles. The outer circle represents the shape of the expanding main wave a moment later and the inner circle, a moment earlier. This idea works in a symmetric way if the central big circle happens to be a contracting wave front—just reverse the words "later" and "earlier."

Animated Applet

For an animated version of Figure 17.3, see Chapter Applet 25 on the CD-ROM.

This means that if we mask out all but one point of the wave front, we still get new waves emanating from that point-source.

We see this in small estuaries, where wave motion from a larger body of water makes its way in through a small opening. Two openings likewise serve as two new sources, which we see near the center of Figure 17.4:

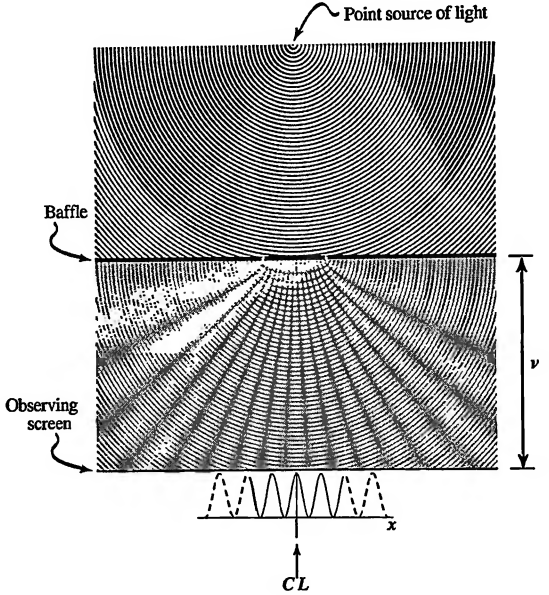


FIGURE 17.4. At the top, a point source of light creates a wave train. When the wave train encounters the baffle with two small openings, we get to see Huygens' Principle in action: the baffle absorbs all the children-waves except those making it through the two small openings. These openings act like the two vibrating needles attached to a tuning fork, mentioned earlier. The bottom half of the figure, resembling the bottom half of Figure 17.2, is illustrated using less light/dark contrast to depict the smaller amplitude (less energetic) waves emanating from the openings. Most of the energy in the top half has been absorbed by the baffle. The graph at the bottom represents the light intensity of the interference fringes.

In this way, a single wave source can be split into two. Since they come from a common source, the two "child" sources vibrate at the same rate as the original, thereby generating the hyperbola fringes shown in the bottom half of Figure 17.4.

Fringes and Beats Magnify

Student. How is all this going to lead to the incredible magnifications you mentioned?

Teacher. A three-word answer is, "Beats can magnify." We see this phenomenon popping up in all sorts of situations. For example, any pilot of a twin engine prop aircraft has heard the beats of one engine against the other. The time interval between successive engine firings is a tiny fraction of a second, too small for the firings to be counted individually. But a small difference in timing between *two* engines running at nearly the same speed is easy to pick up. If we hear one beat per second, we know that if one is firing at N times a second, the other must be firing $N \pm 1$ times a second. Beats magnify the difference enough so that we human beings can sense it.

Philosopher. Piano tuners use the same technique, don't they?

Teacher. Absolutely. If you pluck one string, and a second or two later pluck another string that's a "quarter-tone" higher in pitch, most people can easily hear a difference. Some can hear even an eighth-tone difference—and dogs do considerably better than that. But in the realm of piano turning, you're talking about much less than a quarter or eighth tone. For example, one commonly-used frequency for middle C is about 523 oscillations a second. Now suppose that two of the three middle-C strings of a piano are oscillating 523 times a second, but that the third is out of tune, oscillating 524 times a second. That higher pitched string turns out to be a mere 64th of a semitone above middle C. If you pluck a 523-string and then, a second or two later pluck the 524 string, very few people can hear any difference in pitch. However, overlapping the sounds—sounding the two strings *together*—generates about one beat per second. This is *extremely* easy to hear, and most people react by thinking the piano is out of tune!

> The further you go up the piano keyboard, the more remarkable the power of beats becomes: one octave above middle C, one beat a second corresponds to 128th of a semitone, and at an octave above that, to 256th of a semitone.

One can use the term "octave" to mean *any* 2:1 frequency ratio, even if the waves are not sound, but electromagnetic. In this sense, the spectrum of visible light vibrates in the range of 39.5 to 40.4 octaves above middle C. If the ability to discern differences is remarkable at one or two octaves above middle C, what it can do around 40 octaves is simply spectacular. For example, those rainbow colors you see on a soap bubble are visual beats, in which light reflecting from the inside and outside layers of the bubble interfere. This can serve as an extremely accurate way to measure the thickness of the soap film. As you look at a fresh soap bubble, gravity pulls the soap solution downward, gradually thinning the film; shortly before the bubble gets so thin that it bursts, you'll see an area go colorless or black. The film there has a thickness of about half the wavelength of violet light. The light is destructively interfering, canceling out all the colors visible to the human eye.

Student. How thin is thin?

Teacher. You've seen those little millimeter marks on the centimeter side of a ruler. Subdivide one millimeter into 5000 equal parts—that's about the thickness of a dark, colorless region of a soap bubble.

Philosopher. That is really impressive. I've often seen rainbow colors in bubbles, but never thought of them as revealing small differences in their wall thickness.

Astronomy and Small Angles

Student. I don't see hyperbolas in *any* of these examples!

Teacher. A lot of different astronomical measurements come down to the hyperbola picture in Figure 17.4, or a variant of it. For example, although stars are hurtling through space at typically tens or hundreds of miles a second, our enormous distance from them means that their angular speeds as seen from here are extraordinarily small. With enough time, the angular displacements do build up—the heavens of the ancient Greeks 2500 years ago in fact look quite different now. But measurements taken a month or a year apart typically translate to exceedingly small fractions of a second of arc. Resolving such small angles similarly arises in trying to measure optically the angular distance between closely-spaced binary stars (separated by perhaps only a star diameter or two), or measuring the diameter of a star's disk. These can be daunting challenges. It was Albert Michelson (1852–1931) who demonstrated that one can use optical beats, or interference, to carry out very precise measurements, and in fact he found the diameters of some of Jupiter's moons using this method. Michelson eventually won the 1907 Nobel prize in physics for his pioneering work in optical interferometry.

In the last few decades, enormous strides have been made in applying interferometry to astronomy, and today we are looking forward to a resolution of one *millionth* of a second of arc!

To explain how to harness the power of interference fringes to these tasks, let's contemplate the idea of *diffraction ruling*: extremely close-together parallel lines. Take two transparencies of very thin parallel lines, one drawn at 500 per inch, and the other at, say, 502 per inch, and place one atop the other, with all lines being parallel. You'll see a system of alternating light and dark fringes. These fringes are detecting that small difference between the two rulings, yet the dark and light fringes are each a quarter of an inch wide. Relative to the underlying spacing, this is enormous. The light fringes represent locations where the lines are approximately "in phase" and therefore overlap, while the dark ones are where the lines are approximately "out of phase." Now let's keep one transparency fixed and, perhaps with a screw mechanism, very slowly translate the other. This extremely slow motion changes the locations where the fringes are in or out of phase and as a result, you see the system of fringes parade across the sheets. This is not unlike magnification when using a microscope: if you move the mounted specimen just a tiny bit by hand, it seems to race across the viewing field.

You are in effect witnessing a magnification of either transparency. In moving one transparency a full "grating wavelength"—just $\frac{1}{500}$ of an inch—you've moved one full "fringe wavelength," which in our case is $\frac{1}{2}$ inch. That represents a magnification power of 250.

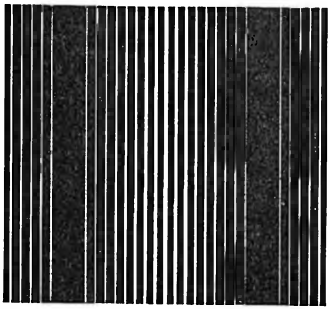


FIGURE 17.5. A 500-line-per-inch ruling looks gray, in analogy to the rapid firing of an airplane's engine cylinders sounding like a "hum." It's when high-frequency waves overlap that you hear or see beats. This picture's individual grating lines have been magnified by about 50 times so we can see them, analogous to slowing down an engine so we can hear individual piston-firings. To enable a fringe or two to appear on this relatively small page, we've chosen a (50-times magnification of) a 525-lines-per-inch grating rather than 502. The lines are approximately in phase at the center, and about 180° out of phase around the middle of the dark bands on either side.

Interactive Applet

For an interactive version of Figure 17.5, see Chapter Applet 26 on the CD-ROM.

Student. How does this apply to astronomy?

Teacher. Suppose that the point source of light at the top of Figure 17.4 is a star. A two-slit baffle, down here on earth, creates two small sources of light, which an appropriate filter can restrict to one wavelength. Faint hyperbolic fringes are created, depicted in the bottom half of the picture. Any particular fringe corresponds to a fixed difference in the radii of Figure 17.1 [p. 342]. On the viewing screen, successive darkest areas, for example, correspond to a difference in radii of one full wavelength.

Now look at Figure 17.6. The point $S = (0, v)$ represents the light source in Figure 17.4. S lies on the vertical line of symmetry of the isosceles triangle SPQ in Figure 17.6, so the difference in their side lengths is zero. Let a second point source $S' = (x, y)$ create a small angle $\Delta\theta$ encountered, for example, in measuring a star's position at two different times, or in measuring the angle subtended on earth by two binary stars. The triangle $S'PQ$ is *not* isosceles, so the difference in side length is no longer zero. To find this difference, we can use differentials.

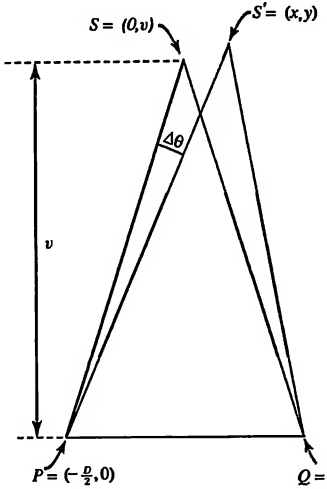


FIGURE 17.6. In this drawing, the left side of triangle $S'PQ$ is longer than its right side. If S' and S are close, then differentials can give a good approximation to the difference in side lengths.

From Figure 17.6, we see that the difference $F(x, y)$ in these lengths is

$$F(x,y) = \sqrt{\left(x+\frac{D}{2}\right)^2 + y^2} - \sqrt{\left(x-\frac{D}{2}\right)^2 + y^2}.$$

Working out $dF = F_x\,dx + F_y\,dy$ and evaluating it at $(0, v)$ leads to

$$\Delta F(0,v) = \frac{D\cdot\Delta x}{\sqrt{\frac{D^2}{4}+v^2}} \approx \frac{D\cdot v\cdot\Delta\theta}{\sqrt{\frac{D^2}{4}+v^2}}. \tag{17.1}$$

The last approximation follows from Figure 17.6, since for a star, the altitude of triangle $S'PQ$ is much greater than its base; therefore $\Delta x \approx v\Delta\theta$. Likewise, v is much greater than D, so that

$$\frac{v}{\sqrt{\frac{D^2}{4}+v^2}} \approx 1,$$

and (17.1) yields

$$\Delta F \approx D\cdot\Delta\theta. \tag{17.2}$$

Let the light's wavelength λ serve as a natural unit of length; the length of ΔF relative to this unit of measurement is therefore approximately $\frac{D \cdot \Delta\theta}{\lambda}$. This is the number of full fringe-widths by which the fringe system has moved away from the original system in Figure 17.4. For example, if $\Delta F = 3$, then as a star moves from S to S' in Figure 17.5, we see the train of fringes move over three "train-cars," or fringe-widths. We can solve for $\Delta\theta$:

$$\Delta\theta \approx \frac{\lambda \cdot \Delta F}{D}. \tag{17.3}$$

Keeping in mind how very far away the star is, and how small $\Delta\theta$ is, we see that this approximation is exceedingly good.

But there's a clinker in (17.3). It's that ΔF may be far, far less than 3. $\Delta\theta$ can be *so* minute that ΔF is but a tiny fraction of a full fringe-length. This translates visually to a small change in fringe intensity—its darkness or lightness at a point. This change can be so slight that even with state-of-the-art photon detectors that pick up very small variations in light-intensity levels, it may not be possible to say just how much the fringe system has moved. And of course knowing this is essential to finding $\Delta\theta$. But there *is* some wiggle room—(17.3) tells us that we can make ΔF larger—and therefore more likely to be measurable—by looking at a smaller wavelength λ and/or by making D larger. We saw a moment ago how making λ smaller increases sensitivity, as we moved in our examples from middle C to one and two octaves higher. Historically, implementing this has presented practical problems, but astronomers have slowly progressed from radio-frequency lengths to X-rays. Astronomers have also found clever ways of effectively increasing D. This has had spectacular results, and makes for great storytelling.

Student. Love it!

A Tale of Progress

Teacher. Figure 17.4 illustrates the basic principle behind creating interference fringes from a single light source, and working with Figure 17.6 shows that from our observation point, an angular change of the light source in Figure 17.4 translates to moving the fringes and thus to measuring small angles. The theory is truly beautiful, but there's this little problem of *reality*...

Philosopher. What do you mean?

Teacher. I actually mean there are *several* problems! Let me list a few:

- Stars don't send us much light to begin with;
- We filter it to make it monochromatic;
- We admit only the tiny amount passing through two narrow slits in the baffle;
- A bit of elementary optics tells us that to see fringes, the observing screen must be at least $\frac{D}{2}$ from the baffle, and that means the separation D can't be too small—which in turn means that the amplitude of the wavelets further decreases before they can begin to interfere;
- The fringes of interest are closer to the center of the screen, meaning that most of the light making it to the screen in Figure 17.4 is wasted.

Student. That's a lot of problems! Wouldn't they make the odds almost impossible?

Teacher. It would seem that way. But in spite of everything, astronomers have succeeded remarkably well. By dint of a series of improvements and conceptual breakthroughs, they are now actually getting close to the maximum of what the theory promises. Let's briefly shake hands with some of these improvements.

Here's a picture showing one way to reduce the amount of wasted light:

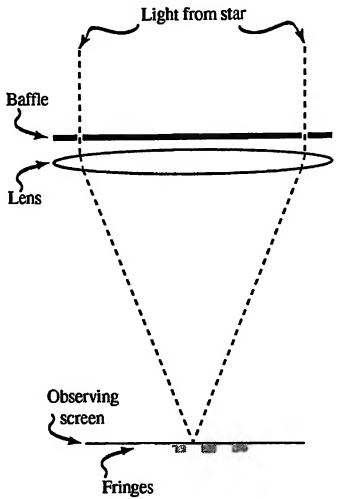


FIGURE 17.7. The converging lens captures nearly all the light passing through the baffle and focuses the split beam toward the center of the screen, where one sees interfering beams creating light and dark fringes.

The idea of this setup was used by Michelson in some of his early experiments. The lens takes nearly all the light and focuses it toward the center of the observation screen. Result: brighter, clearer fringes. Over a large telescope's 12-inch diameter lens (the lens facing the heavens), Michelson placed a two-slit baffle with an adjustable distance D between the slits. In those days there was no precise, quantitative way to measure fringe intensity such as a photon detector, so in using his setup to measure the angle between two binary stars, he instead adjusted D until the fringes seemed to disappear completely—which happens when the fringes from the two stars are out of phase. With everything on the right-hand side of (17.3) known, he could then calculate $\Delta\theta$.

Making D something like 12 inches was good, but, all else being equal, the larger D is, the better. By being creative with mirrors, Michelson had by 1920 constructed a setup with twenty times that aperture size.

Figure 17.8 shows the concept:

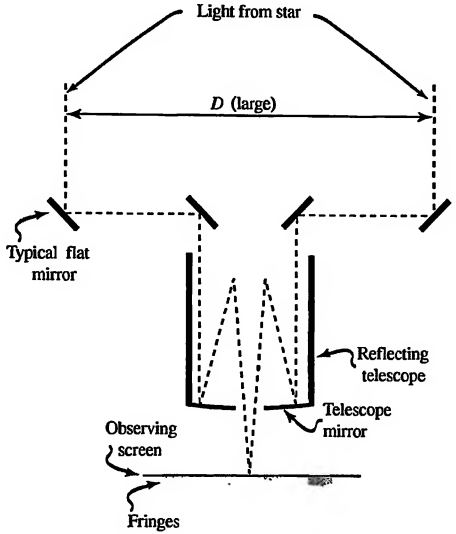


FIGURE 17.8. Here, instead of a converging lens, mirrors bring the two light beams closer together before being fed into a reflecting telescope.

With this apparatus, he was able to attain several times the resolution of his earlier setups, and by 1930, he'd built a version of Figure 17.8 having a *fifty* foot separation. Heavy mirrors were solidly affixed to nine tons of steel girder work built in front of a 40-inch telescope. This proved to press the limits of what was then possible: the atmospheric turbulence that causes stars to twinkle made the interference fringes jitter around, causing errors in excess of 15 percent.

Radio Waves

Teacher. The next chapter of our story proceeds in a direction opposite from what you'd expect. The formula in (17.3) suggests that working with ever shorter wavelengths should be the way to go. But within the visible spectrum, the shorter the wavelength, the more the atmosphere scatters light, causing stars to "twinkle." Blue light, at the short end of the wavelength rainbow, is the most easily scattered, and all that scattered blue light from the sun is what makes the sky blue. We see the reverse of this in the setting sun: most of its

blue light, on the way toward our eyes, has been scattered by many miles of atmospheric dust and moisture droplets near the horizon, making the longer reddish light waves the "survivors." We see those survivors as the familiar orange-red ball at sunset.

If long-lengthed red light is such a survivor, then what about even longer wavelengths such as, say, radio waves? In the 1930s, radio waves became increasingly used to transmit transoceanic calls. One major annoyance was the considerable interference and static that seemed to originate from somewhere in outer space. In 1932, Bell Telephone Labs wanted to determine just what was the source of all this interference. Bell Labs asked one of its engineers, Karl Jansky (1905–1950), to investigate. Jansky built a large radio-receiving antenna to help determine what part of the sky the offending noise was coming from. The sun seemed to be the likely culprit—after all, it's the source of almost all the electromagnetic energy we receive. To Jansky's surprise, the noise seemed to come instead from our own galaxy, the Milky Way. It wasn't possible to determine specific locations for the signals: radio waves have about the same length as sound waves, and the ability to pinpoint sound is limited—its waves are about a yard long.

> Our eyes are far better able to pinpoint direction than our ears, because light waves are so much shorter—the shorter the wavelength, the better the theoretical resolution. No one can discern as distinct two sound sources placed like two light bulbs five feet apart three miles away! One can do that with light waves, but not with sound or radio waves.

Now perhaps it was because the data was so approximate. Or perhaps it was because "only noise" was being measured. But one thing was certain: hardly *anyone* took much notice of the new discovery. Years later, in 1940, a radio engineer and avid ham radio operator, Grote Reber (1911–2002), chanced upon a story in *Popular Astronomy* about Jansky's work on the cosmic hiss. *What was going on*? Reber was an eccentric, highly ingenious character, always on the lookout for fresh challenges. Some while before he saw the article, he'd pretty much concluded that in radio, there were no more worlds to conquer. And then Jansky's work seemingly appeared from nowhere! Here, it seemed to Reber, was an extremely promising idea—a surprising result, a new surface that had barely been scratched. He'd found the ideal challenge.

His keen engineering instinct led him to a parabolic dish-shaped antenna, and with his own very limited funds, he built a 30-foot receiver in his mother's back yard. That was the world's first radio dish, and it already was a great improvement over Jansky's antenna. With his dish, Reber made "radio maps" of the heavens. Though accurate to only a few degrees—and therefore far less precise than the human eye—he was mapping something new in the sky.

After WW II ended, a number of scientists who had had a hand in developing radar cast about for new challenges, and Reber's work caught their attention. Reber's dish, though impressive, lacked the resolving power to enable astronomers to say, "Look! This radio-frequency source is originating from such-and-such a star." Calculations showed that a dish *that* good would have to be more than a half-mile wide!

But there was a way out. Think back to the focus-to-focus property of an ellipse. If some (or even most) of the ellipse is missing, then instead of *all* the light from the source making its way to the other focus, only some of it does. The rest of the light simply escapes through the holes or gaps in our incomplete ellipse. But whatever light *does* impinge on the incomplete ellipse goes to the other focus.

Philosopher. I see that. When we use a reflector to focus the sun's rays to start a fire, we're certainly not capturing all the sun's energy—just a fantastically minute portion of it.

Teacher. Great illustration! So instead of building an entire half-mile-wide dish, one could instead build a little of it here, a little of it there. That way, although you make a very incomplete dish, if you have portions separated by a half mile, you'll still get the resolving power of the entire dish. You don't get as much *intensity*, but you do get greatly improved *resolution*.

Student. But just a minute ago you said that Michelson's fifty-foot arrangement pressed to the limit—that the atmosphere made the fringes jitter around. If fifty feet do that, what would a half mile do?!

Teacher. The big difference is that "radio waves don't twinkle." That is, they're much less affected by atmospheric vagaries—air turbulence, moisture droplets, dust and the like—than any wavelength of visible light. That means the ability of fringes to magnify isn't corrupted by turbulence, and in fact the distance can be made thousands of miles! This basic fact opened the way to tremendously improving the resolving power of "radio telescopes." And successes came: by 1949, resolving power increased to 10 minutes of arc; by 1951, to 1 minute—about as good as normal human eyesight.

Student. So why not just look at the sky with your eyes? Why spend millions of dollars to do the same thing with radio waves?

Teacher. One good reason is that the universe can look very different through "radio eyes." There are astrophysical processes going on that we can't pick up in the range of visible light, and these often tell a surprising and informative story at longer wavelengths. As an example, there are legions of "could-have-been" stars—ones that didn't have enough mass to collapse sufficiently to create a nuclear furnace. So they're hot, but not glowing. Being so far away, we see nothing of any of their reflected light, with even high-powered conventional telescopes. Often, however, they emit a profusion of longer wavelengths, and often radio telescopes pick them out like beacons in the sky. Radio telescopes have had other successes, too. For example, radio receivers were directly responsible for discovering tremendously powerful "radio galaxies," quasars, and black holes.

Radio telescopes allow us to look far deeper into the center of our own galaxy than visible light permits. This is because interstellar dust scatters visible light, but radio waves pass on through. Most of what we know of the center of our Milky Way Galaxy has come courtesy of radio waves—it's hard to overstate their importance.

Student. That sounds impressive. Now how does this tie in with the fringes and tremendous magnification that we saw in the visible spectrum? Are radio waves going to outdo what was accomplished by Michelson and others using ordinary light?

Teacher. In some ways, yes. Just as optical interferometry followed on the heels of conventional optical telescopes, radio interferometry followed single dishes. A good example is the VLA—Very Large Array—located about 50 miles west of Sorroco, New Mexico. This is a group of 27 large, 240-ton dishes arranged in a large Y-shape and mounted on railroad tracks so their configuration can be adjusted. The dishes can achieve an effective diameter of 26 miles, creating a very large D in (17.3). Signals from each dish are sent via cable to a central processing center, and are electronically combined. Depending on the wavelength—it can handle wavelengths from about a yard to a quarter inch—and with the large overall diameter of the Y-shape, the VLA can resolve to 0.04 seconds of arc. That beats than any single ground-based optical telescope!

The separation D in (17.3) can be increased yet further: radio disks from around the world, and even in space, can combine signals to achieve unprecedented resolution. The VLBI—Very Long Baseline Interferometer—combines signals from satellite and earth-bound disks to make D nearly three times the earth's diameter! Pictures from it are incredible.

Student. Do you really mean "pictures"? That would be exciting. I thought we were only looking at *fringes*.

Teacher. Yes, I mean actual pictures. Consider again Figure 17.4 [p. 346]. That single wave source creates a diffraction pattern on the observing screen. That is, the top half of the picture leads to the bottom half. But this order is reversible in the sense that starting from the bottom half of the picture, you can (mathematically) recreate the top half—and that's true even for multiple light sources.

> It's an oversimplification, but the spirit in reversing the order of the top and bottom halves of Figure 17.4 on p. 346 is, in technical jargon, like taking a "Fourier transform." In our setup, point-like information from the sources ends up as fringe-information on the observation screen, this information being distributed or wavelike in character. Going from the bottom half to the top—reconstructing the original image from the fringes—corresponds to taking the inverse Fourier transform. This same idea is used in making MRIs or CAT scans, which reconstruct 3-D images from wave data. This very general and widely-used process has been made much more practicable by Fast Fourier Transforms (FFT), introduced in 1965.

The resolution of the VLBI can be more than a thousand times that of the best optical telescopes on earth.

Philosopher. It just occurred to me that although using a small wavelength is difficult on earth due to scattering, we can put instruments in space and use much shorter wavelengths there.

Modern software has made image recreation from wave data fast and automatic. Astronomers can choose color schemes, and utilize bells and whistles like "image noise reduction" and "adaptive smoothing" to create informative and detailed pictures of the center of our galaxy, of matter streaming toward a black hole, and of galaxies near the edge of our universe.

Teacher. The Chandra X-ray telescope is an excellent example of just that. Put into operation in 1999, it has, for example, provided unprecedented views of the center of our galaxy; X-rays, like radio waves, are able to penetrate through its dusty, "smog-like" center. Such detailed views of the busy center are basic to increasing our understanding of galaxies all over the universe.

Student. If we go into space, why not use several observatories to do in space what VLA does in New Mexico?

Teacher. In fact, NASA is planning to launch, in 2010, a fleet of some 33 small spacecraft, each with an X-ray telescope. These will fly in precise formation to make a virtual 300-mile-wide X-ray dish. This is expected to produce images *three million* times sharper than the Chandra X-ray observatory. These are indeed exciting times for using Figure 17.4 to help unlock the secrets of the universe!

Chapter Summary

The hyperbola construction in Figure 17.1 models a natural phenomenon that scientists have been able to exploit to make breathtakingly precise measurements. Analyzing fringes or "beats" generated when wave trains interfere—interferometry—has a wide range of applications. A two-point, commom-frequency wave source generates a pattern of interference bands that is naturally described by the confocal system of conics appearing in Figure 17.2 [p. 343]. The closer together the two sources, the wider the hyperbola-branch bands, and this phenomenon serves as a natural "magnifier"; it can be used as an extremely powerful way of bringing the heavens closer to us. Although the tremendous potential of such magnifying power is rife with practical difficulties, astronomers and physicists have persisted through the decades, and today we stand at the crest of crucial advances in the art of utilizing this phenomenon. This chapter includes enough detail for the reader to see that those broad bands are in fact hyperbola branches. In Exercise 5, we see that the orthogonal-appearing bands in Figure 17.2 form ellipses in a confocal system of conics.

Exercises

1. It is known that bees can see ultraviolet light, while we can't. That means the world looks different to them—for example, every buttercup has an ultraviolet-reflecting ring around its center, and bees see this UV halo. Does this mean that bees could better judge when a soap bubble is about to burst?

2. In Figure 17.5 on p. 349, suppose we'd used a "502" grating rather than a "525" grating. At the 50-fold magnification used in this figure, about how far apart would neighboring dark fringes be?

3. In Figure 17.5, suppose the number n of vertical bars per unit horizontal length is large, and let m be an integer with small absolute value. Find a magnification formula for the fringes created when vertical gratings of n and $n + m$ lines per unit horizontal length are superimposed.

4. Can you create a family of *parabolas* by combining a circular grating with a linear one? (Intuitively: can parabolic interference fringes be created when ripples emanating from oscillations centered at the origin meet ripples emanating from oscillations centered at a point at infinity?)

5. About those fainter confocal ellipses noticed in Figure 17.2 on p. 343: how do we know they're actually ellipses? To see why, freeze time, index the crest-circles centered at the left point by integers m, and index the crest-circles centered at the right point by n. For our hyperbola branches, as m increases by 1, so does n, and that means $m - n$ is constant. A little thought shows that instead of using only the difference, *any* integral linear combination $\alpha m + \beta n$ defines a path—in particular, $m + n = k$ (k, a constant integer) defines paths of maximum oscillation in the same way the difference does. To get the equations of these "sum" paths, let the left circles have equations

$$(x + 1)^2 + y^2 = m^2$$

and the right ones,

$$(x - 1)^2 + y^2 = n^2.$$

Solve for m and n, and assume $m + n = k$, with k a constant greater than 2. This leads to

$$\sqrt{(x - 1)^2 + y^2} + \sqrt{(x + 1)^2 + y^2} = k,$$

just a stakes-and-string condition for a real ellipse. Show that the equation of this ellipse is

$$\frac{x^2}{\left(\frac{k}{2}\right)^2} + \frac{y^2}{\left(\frac{k}{2}\right)^2 - 1} = 1.$$

As k varies, the curves fill in a system of confocal ellipses.

Appendix 1
Some Conics Formulas

Introduction

This appendix contains a collection of useful formulas about conics, together with some general comments. Conics arise in a variety of ways, ranging from geometric to astrophysical to algebraic. Correspondingly, their equations have a considerable range of forms. But no matter how a general conic is specified, it always has a net of five independent degrees of freedom. These can range from geometric parameters specifying a cone and slicing plane, to physical ones giving a satellite's position and velocity at some instant, to purely algebraic parameters, such as the coefficients of the quadratic equation $Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0$. In each case there are formulas for the conic's vitai statistics, like center, major and minor axes, eccentricity, focal length, area, and so on.

In practice, one often uses more than five parameters in specifying a conic. For example the quadratic equation $Ax^2 + \cdots + F = 0$ has six coefficients, not five. But among these coefficients there is one degree of redundancy: the six coefficients are not uniquely determined, because the locus is unchanged when we multiply the equation through by a nonzero constant. There are thus $6-1 = 5$ degrees of freedom. Another example is choosing five distinct points in the plane: except in certain degenerate cases, there's a unique conic passing through these points. Since each point has two degrees of freedom—its x- and y-coordinates, for example—we've actually invoked a total of 10 parameters. On the other hand, each of the five points yields one degree of redundancy, or slack, because the point can move along that one-dimensional conic without changing the curve. Thus the number of independent degrees of freedom is $10 - 5 = 5$.

Specifying the location of a conic requires two degrees of freedom. One often fixes a point like the center or a focus, and then there remain three degrees of freedom. A typical example is the quadratic $Ax^2 + Bxy + Cy^2 = 1$, where the center is at the origin. In this case, lurking in the three pieces of *algebraic* data A, B, C are three pieces of *geometric* data: shape, size and orientation. These might be given by the eccentricity ε, length a of the semi-major axis, and that axis' inclination ϕ to the horizontal. Since A, B, C determine the conic, there should be formulas in terms of A, B, C for all its vital stats. We supply formulas of this kind for eight different methods of specifying conics.

The formulas listed below sometimes produce imaginary values, and these imaginaries can pop up in dissimilar-appearing formulas. Nonetheless, there is great consistency. For example, formulas for area always become imaginary exactly when the conic is a hyperbola. Look, say, at the cone/plane list of formulas below. The area there is

$$\mathbf{A} = \frac{\pi H^2}{(1 - \tan^2 \psi)^{3/2}},$$

and this is imaginary when $\tan^2 \psi > 1$—which just says the intersecting plane's angle ψ is large enough to make the plane intersect the cone in a hyperbola. Likewise, area in the polar equations list is

$$\mathbf{A} = \frac{\pi p^2}{(1 - \varepsilon^2)^{3/2}},$$

and that is imaginary when $\varepsilon > 1$—precisely the condition for a hyperbola, illustrated in Figure 7.2 on p. 107. In the stakes-and-string list, $\mathbf{A} = \pi L\sqrt{L^2 - \mathfrak{f}^2 - \mathfrak{g}^2}$ is imaginary when $L^2 < \mathfrak{f}^2 + \mathfrak{g}^2$, and this inequality, expressed in terms of a and b, becomes

$$0 < a^2 < (\text{focal distance})^2 = a^2 - b^2,$$

which implies $b^2 < 0$. Of course, in

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1,$$

a positive value or a^2 and a negative one for b^2 specifies a hyperbola. Or in the quadratic forms list,

$$\mathbf{A} = \frac{2\pi I}{\sqrt{B^2 - 4AC}},$$

and this is imaginary whenever the form's discriminant $B^2 - 4AC$ is positive. Of course, the discriminant $B^2 - 4AC$ famously discriminates, telling us that $Ax^2 + Bxy + Cy^2 = 1$ defines a hyperbola exactly when $B^2 - 4AC$ is positive. And in the "astrophysics" list, we have

$$\mathbf{A} = \frac{\pi h}{(-8E)^{3/2}}.$$

Positive total energy ($E > 0$) makes the area imaginary, and it happens that only when $E > 0$ does an asteroid possess the energy to escape the grip of its star's gravitational field, allowing it to travel along a hyperbola branch.

The Formulas

In the formulas that follow

- a denotes the semi-major axis;
- b denotes the semi-minor axis;
- (h, k) denotes the conic's center;

$\|F\|$ denotes the distance (separation) from the center of the conic to a focus;

$\|D\|$ denotes the distance (separation) from the center of the conic to a directrix;

ε denotes eccentricity;

p denotes the semi-latus rectum;

ϕ denotes the angle of tilt of the semi-major axis to the horizontal;

A denotes area;

κ_a denotes the curvature at one end of the major axis;

κ_b denotes the curvature at one end of the minor axis;

Here are eight different approaches, with some explanation and perspective for each.

1. Plane Intersecting a 45° Cone. The equation $z^2 = x^2 + y^2$ defines a standard 45° cone in $\mathbb{R}^3$. Any conic in the (x, y)-plane is then the vertical projection of this cone's intersection with some plane in $\mathbb{R}^3$. Each of these conics, when nondegenerate, has a focus at the origin. The conic's location thus determined, there remain three degrees of freedom, and since the cone is fixed, these degrees of freedom must reside in the slicing plane. Obligingly, a general plane in 3-space has exactly three degrees of freedom. As shown in Figure A1.1, we may take these parameters to be: the plane's intercept H with the z-axis; and the polar and azimuthal angles ψ and ϕ of a line normal to the plane.

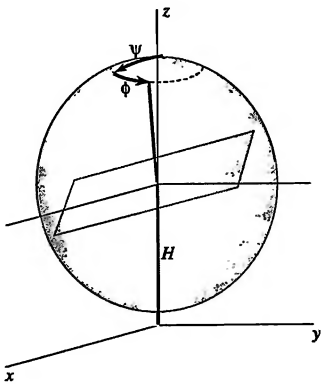


FIGURE A1.1. This shows how an arbitrary plane in 3-space is determined by the three parameters H, ψ and ϕ: shift the (x, y)-plane up by H; then rotate it by tilting its normal an amount ψ in the (x, z)-plane; then follow with a rotation ϕ about the z-axis.

The parameters H, ψ and ϕ determine a plane, and therefore they determine size, shape and tilt of the conic's projection into the (x, y)-plane. This is illustrated in Figure A1.2.

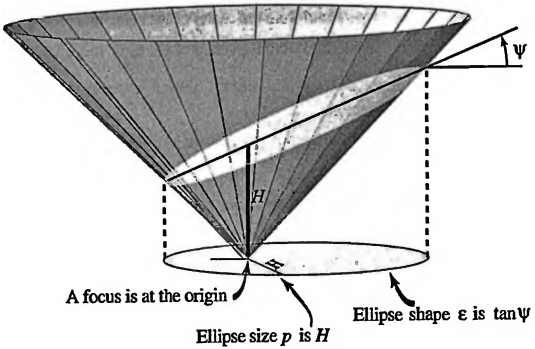


FIGURE A1.2. The three parameters H, ψ and ϕ specify a slicing plane, which in this illustration determines the size, shape and tilt of an ellipse; we've taken $\phi = 0$. The cone's axis passes through a focus. When the slicing plane is rotated by ϕ around this axis, the projected ellipse in the (x, y)-plane rotates by ϕ, too.

H determines size. Increasing $H > 0$ translates the plane vertically upwards and corresponds to "zooming in" on the conic in the (x, y)-plane. The magnification is linear—for example, doubling H doubles the linear dimensions of the conic. Decreasing H corresponds to zooming out, so when H is very small, we see a conic viewed from a great distance. From there, an ellipse looks like a point, a parabola looks like a ray, and a hyperbola looks like its asymptotes.

ψ determines shape. With ϕ and $H > 0$ fixed, let ψ increase from 0. The projected conic in the (x, y)-plane begins as a circle and then elongates, running through all possible ellipse shapes as ψ varies in the interval $0 < \psi < \frac{\pi}{4}$. We get a parabola when $\psi = \frac{\pi}{4}$, and all possible hyperbola shapes when $\frac{\pi}{4} < \psi < \frac{\pi}{2}$. When $\psi = \frac{\pi}{2}$, the projected conic in the (x, y)-plane degenerates to a double line through the plane's origin. Of course as ψ continues to increase, the scenario reverses, repeats, reverses ..., during successive $\frac{\pi}{4}$ increments in ψ. A natural question arises: *why use vertical projections rather than the slices themselves*? The answer is that the unprojected slices don't run through all possible conic shapes! To see why, look at Figure 7.2 on p. 107, illustrating the correspondence between shape and eccentricity. It happens that for the unprojected conics, $\varepsilon = \sqrt{2}\sin\psi$. This means the maximum value of ε is $\sqrt{2}$, attained when $\psi = \frac{\pi}{2}$. We see from Figure 7.2 that as $\epsilon \to \infty$ (that is, as ψ approaches $\frac{\pi}{2}$), the hyperbola branches straighten out, in the limit looking like a double line.

ϕ determines tilt. Keeping ψ and H fixed and varying ϕ causes the conic to rotate an amount ϕ about its focus at the origin.

The equation of the projected conic in Figure A1.2 can be written in polar coordinates, and in this form the three parameters H, ψ and ϕ all appear explicitly:

$$r = \frac{H}{1 - \tan\psi \cos(\theta - \phi)}. \tag{A1.1}$$

Since H, ψ and ϕ completely determine all conics with focus at the origin, other measurements can be expressed in terms of them. Here is a list.

$$a = \frac{H}{1 - \tan^2\psi} \qquad b = \frac{H}{\sqrt{1 - \tan^2\psi}}$$

$$h = \frac{H \cdot \tan\psi \cdot \cos\phi}{1 - \tan^2\psi} \qquad k = \frac{H \cdot \tan\psi \cdot \sin\phi}{1 - \tan^2\psi}$$

$$\|F\| = \frac{H \cdot \tan\psi}{1 - \tan^2\psi} \qquad \|D\| = \frac{H \cdot \cot\psi}{1 - \tan^2\psi}$$

$$\varepsilon = \tan\psi \qquad p = H \quad [p \text{ doesn't depend on } \psi!]$$

$$\kappa_a = \frac{1}{H} \qquad \kappa_b = \frac{(1 - \tan^2\psi)^{3/2}}{H}$$

$$A = \frac{\pi \cdot H^2}{(1 - \tan^2\psi)^{3/2}}$$

Observe that ε is negative exactly when ψ is. Although we could drop the negative sign, it provides information on which focus is at the origin: when $\psi > 0$, the origin is the ellipse's left focus or the hyperbola's right focus; when $\psi < 0$, the origin is the ellipse's right focus or the hyperbola's left focus.

Interactive Applet

For an interactive applet showing how ψ, ϕ and H determine conics with a focus at the origin, with on-the-fly values of the above formulas, see Appendix Applet 1a on the CD-ROM.

2. Polar Equations. The relation between the cone/plane method and the equation in polar coordinates is very close, as the equation in (A1.1) shows. The polar equation is essentially the algebraic counterpart of the geometry in Figure A1.2, and provides yet another example of how mathematical "brothers" can be discovered centuries apart: plane sections of a cone date from ancient Greece, while polar coordinates had to wait until at least Newton—historians surmise he may have stumbled across them—and perhaps even until 1691, when we know for sure that Jakob Bernoulli (1654–1705) used them.

There are actually three natural versions of the polar equation of a conic. The first two in our list correspond to placing the origin at one focus or the other, each modeling a planetary orbit about a massive star in the case of an ellipse. In the third equation, the origin is at the conic's center.

The first two polar equations are:

$$r = \frac{p}{1 - \varepsilon \cdot \cos(\theta - \phi)}; \tag{A1.2}$$

$$r = \frac{p}{1 + \varepsilon \cdot \cos(\theta - \phi)}. \tag{A1.3}$$

The third equation, with origin at the conic's center, is:

$$r^2 = \frac{p^2}{(1 - \varepsilon^2)(1 - \varepsilon^2 \cdot \cos^2(\theta - \phi))}. \tag{A1.4}$$

Adding π to ϕ interchanges (A1.2) and (A1.3), and therefore amounts to shifting the origin from one focus to the other. When $\phi = 0$, the left focus is at the origin in (A1.2), and the right focus is at the origin in (A1.3). A parabola has one focus at infinity, so the phrase "parabola center" is not defined; notice how, appropriately, the formula in (A1.4) breaks down in this case of $\varepsilon = 1$.

In each case the center or a focus of the conic is fixed at the origin, so there remain only three degrees of freedom: size is given by the semi-latus-rectum p; shape is determined by eccentricity ε; the tilt is ϕ. The formulas below are for

$$r = \frac{p}{1 - \varepsilon \cdot \cos(\theta - \phi)}.$$

Actually, except for the center (h, k), the formulas are valid for all three equations (A1.2)–(A1.4). This implies that for given p, ε and ϕ, the three conics are just translates of each other.

$$a = \frac{p}{1 - \varepsilon^2} \qquad b = \frac{p}{\sqrt{1 - \varepsilon^2}}$$

$$h = \frac{\varepsilon \cdot p \cdot \cos\phi}{1 - \varepsilon^2} \qquad k = \frac{\varepsilon \cdot p \cdot \sin\phi}{1 - \varepsilon^2}$$

$$\|F\| = \frac{\varepsilon \cdot p}{1 - \varepsilon^2} \qquad \|D\| = \frac{p}{\varepsilon \cdot (1 - \varepsilon^2)}$$

$$\kappa_a = \frac{1}{p} \quad [\kappa_a \text{ doesn't depend on } \varepsilon] \qquad \kappa_b = \frac{(1 - \varepsilon^2)^{3/2}}{p}$$

$$\mathbf{A} = \frac{\pi p^2}{(1 - \varepsilon^2)^{3/2}}$$

Interactive Applet

For an interactive applet showing how p, ε and ϕ determine conics with focus at the origin, and giving on-the-fly values of the above formulas, see Appendix Applet 2a on the CD-ROM.

3. Focus, Directrix, Eccentricity This ancient definition can be traced back as far as Pappus (ca. 290–ca. 350). This approach chooses a line D and a point F not on D; the conic is then the set of all points P for which

$$\frac{\text{distance}(P, F)}{\text{distance}(P, D)} = \text{constant}.$$

Here D is the directrix, F is a focus, and the constant is the eccentricity ε.

- ε is a sort of geometric cousin to the discriminant (which is more algebraic), in that each is a shape-parameter. The conic is an ellipse, parabola or hyperbola when ε is, respectively, < 1, $= 1$ or > 1.
- The distance Δ from point F to line D is a "zoom factor": the larger Δ, the more close-up the view.
- The tilt from the vertical of the directrix is the tilt from the horizontal of the conic.

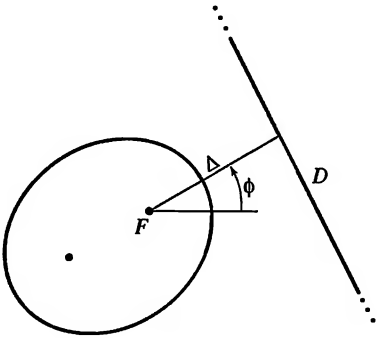


FIGURE A1.3. The five degrees of freedom of a general conic appear in yet another guise: here, two of them pin down a focus, which we've placed at the origin. Two more determine the directrix (Δ and ϕ). The eccentricity ε represents the fifth degree of freedom.

Notice how Figure A1.3 has a polar-coordinates look to it. In fact, a straightforward calculation shows that $p = \varepsilon\Delta$. One can then substitute this for p in the polar list to get a focus-directrix-eccentricity list:

$$a = \frac{\varepsilon\Delta}{1-\varepsilon^2} \qquad b = \frac{\varepsilon\Delta}{\sqrt{1-\varepsilon^2}}$$

$$h = \frac{\varepsilon^2\Delta \cdot \cos\phi}{\varepsilon^2 - 1} \qquad k = \frac{\varepsilon^2\Delta \cdot \sin\phi}{\varepsilon^2 - 1}$$

$$\|F\| = \frac{\varepsilon^2\Delta}{1-\varepsilon^2} = a\varepsilon = \frac{b^2}{\Delta} \qquad \|D\| = \frac{\Delta}{1-\varepsilon^2} = \frac{a}{\varepsilon} = \frac{b^2}{\varepsilon^2\Delta}$$

$$\kappa_a = \frac{1}{\varepsilon\Delta} \qquad \kappa_b = \frac{(1-\varepsilon^2)^{3/2}}{\varepsilon\Delta}$$

$$p = \varepsilon\Delta \qquad \mathbf{A} = \frac{\pi\varepsilon^2\Delta^2}{(1-\varepsilon^2)^{3/2}}$$

Interactive Applet

For an interactive demonstration of how Δ, ε and ϕ determine conics, with on-the-fly values of the above formulas, see Appendix Applet 3a on the CD-ROM.

4. Stakes-and-String Definition of a Conic. Here we fix two foci in the plane, say $F_1 = (-\mathfrak{f}, -\mathfrak{g})$ and $F_2 = (\mathfrak{f}, \mathfrak{g})$, and select a string length $2L$. The set of points whose distances from F_1 and F_2 add to $2L$ forms a conic centered at the origin. When $L > \sqrt{\mathfrak{f}^2 + \mathfrak{g}^2}$, this conic is a real ellipse. Though we state the formulas below in terms of the parameters $\mathfrak{f}$, $\mathfrak{g}$ and L, these formulas immediately extend to the more general five-parameter case of arbitrary foci $F_1 = (\mathfrak{f}_1, \mathfrak{g}_1)$ and $F_2 = (\mathfrak{f}_2, \mathfrak{g}_2)$: simply replace $\mathfrak{f}$ by $\mathfrak{f}_2 - \mathfrak{f}_1$ and $\mathfrak{g}$ by $\mathfrak{g}_2 - \mathfrak{g}_1$. The conic has center

$$\left(\frac{\mathfrak{f}_1 + \mathfrak{f}_2}{2}, \frac{\mathfrak{g}_1 + \mathfrak{g}_2}{2}\right).$$

$$a = L \qquad b = \sqrt{L^2 - \mathfrak{f}^2 - \mathfrak{g}^2}$$

$$\|F\| = \sqrt{\mathfrak{f}^2 + \mathfrak{g}^2} \qquad \|D\| = \frac{L^2}{\sqrt{\mathfrak{f}^2 + \mathfrak{g}^2}}$$

$$\varepsilon = \frac{\sqrt{\mathfrak{f}^2 + \mathfrak{g}^2}}{L} \qquad p = \frac{L^2 - \mathfrak{f}^2 - \mathfrak{g}^2}{L}$$

$$\phi = \operatorname{Arctan}\left(\frac{\mathfrak{g}}{\mathfrak{f}}\right) \qquad \mathbf{A} = \pi L\sqrt{L^2 - \mathfrak{f}^2 - \mathfrak{g}^2}$$

$$\kappa_a = \frac{\mathrm{L}}{L^2 - \mathfrak{f}^2 - \mathfrak{g}^2} \qquad \kappa_b = \frac{\sqrt{L^2 - \mathfrak{f}^2 - \mathfrak{g}^2}}{L^2}$$

Interactive Applet

For an interactive look at the ellipse determined by total string length and position of stakes, with on-the-fly values of the above formulas, see Appendix Applet 4a on the CD-ROM.

5. Parametric Equations of a Conic. An ellipse with center (h, k), semi-axes a, b and tilt ϕ has parametric equations

$$\begin{aligned} x &= a\cos t \cdot \cos\phi - b\sin t \cdot \sin\phi + h; \\ y &= a\cos t \cdot \sin\phi + b\sin t \cdot \cos\phi + k. \end{aligned} \tag{A1.5}$$

An ellipse in $\mathbb{R}^2$ is traced out whenever t increases by 2π.

These equations are remarkable in that each of the five parameters h, k, a, b, ϕ has direct geometric meaning. When $h = k = \phi = 0$, these equations reduce to the familiar

$$x = a\cos t, y = b\sin t. \tag{A1.6}$$

In fact, (A1.5) can be derived by applying a rotation to (x, y) in (A1.6) to create tilt, followed by the translation $(x, y) \to (x - h, y - k)$, which moves the conic's center to (h, k). [See the "action-reaction principle" on p. 144.]

Here are some formulas involving a, b and ϕ.

$$\|F\| = \sqrt{a^2 - b^2} \qquad \|D\| = \frac{a^2}{\sqrt{a^2 - b^2}}$$

$$\varepsilon = \frac{\sqrt{a^2 - b^2}}{a} \qquad p = \frac{b^2}{a}$$

$$\kappa_a = \frac{a}{b^2} \qquad \kappa_b = \frac{b}{a^2}$$

$$\mathbf{A} = \pi ab$$

Interactive Applet

Appendix Applet 5a on the CD-ROM gives an interactive look at the ellipse with tilt ϕ and semi-axes a, b, with on-the-fly values of the above formulas.

6. Conics Defined by a Quadratic Equation. A general conic may be defined as the locus of the quadratic equation

$$Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0. \tag{A1.7}$$

Although the coefficients could be assumed complex, for our purposes we take them to be real. The conic has a center at the origin if and only if $D = E = 0$, and we assume this to reduce clutter in our formulas. In fact, to keep our formulas as simple as possible, we assume our conic has the form

$$Ax^2 + Bxy + Cy^2 = 1, \tag{A1.8}$$

that $B^2 - 4AC \neq 0$, and that not both A, C are negative. These imply that the conic is nondegenerate. When $B = 0$ in (A1.8), the principal axes are horizontal and vertical. More generally, an equation for the union of the two principal axes is

$$Bx^2 + 2(C - A)xy - By^2 = 0; \tag{A1.9}$$

this follows from "eigen-arguments." When $B = 0$ and $A \neq C$, (A1.9) reduces to $xy = 0$, an equation for the union of the x- and y-axes. Notice that if $B = 0$ and $A = C$, then (A1.9) reduces to $0 = 0$; geometrically, the conic is a circle and principal axes are not uniquely determined.

If $B = 0$ in (A1.7), the equation may be written in "standard form":

$$\frac{(x-h)^2}{a^2} + \frac{(y-k)^2}{b^2} = 1, \tag{A1.10}$$

with the usual understanding that one of a, b may be pure imaginary (corresponding to a real hyperbola). Geometrically, choosing $B = 0$ makes the tilt angle 0, leaving four free parameters. This standard form is deservedly popular, since not only is it simple, but all four parameters have direct geometric meaning: a and b are horizontal and vertical stretch factors applied to the unit circle centered at the origin; and h and k are horizontal and vertical shifts of the center.

One can extend the standard form

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

to include the tilt ϕ as well. If the a-axis is rotated through ϕ from the horizontal, then

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

becomes

$$\left(\frac{\cos^2\phi}{a^2} + \frac{\sin^2\phi}{b^2}\right)x^2 + \sin^2\phi\left(\frac{1}{a^2} - \frac{1}{b^2}\right)xy + \left(\frac{\sin^2\phi}{a^2} + \frac{\cos^2\phi}{b^2}\right)y^2 = 1. \tag{A1.11}$$

Replacing x by $(x-h)$ and y by $(y-k)$ of course shifts the center from $(0,0)$ to (h,k).

Let us now assume the conic has been put in the form (A1.8). Then $(h,k) = (0,0)$, and one can express additional geometric information in terms of A, B, C. There are two cases: $A \le C$ and $A \ge C$, and one reads the formulas by respectively choosing (in the usual way) $+$ or $-$ in $\pm$, and $-$ or $+$ in $\mp$. Here are the formulas.

$$a = \sqrt{\frac{2}{A + C \mp \sqrt{A^2 + B^2 + C^2 - 2AC}}}$$

$$b = \sqrt{\frac{2}{A + C \pm \sqrt{A^2 + B^2 + C^2 - 2AC}}}$$

$$\|F\| = \frac{2 \cdot \sqrt[4]{A^2 + B^2 + C^2 - 2AC}}{\sqrt{4AC - B^2}}$$

$$\|D\| = \frac{\sqrt{4AC - B^2}}{(A + C \mp \sqrt{A^2 + B^2 + C^2 - 2AC}) \cdot \sqrt[4]{A^2 + B^2 + C^2 - 2AC}}$$

$$\varepsilon = \sqrt{\frac{2\sqrt{A^2 + B^2 + C^2 - 2AC}}{A + C \pm \sqrt{A^2 + B^2 + C^2 - 2AC}}}$$

$$p = \frac{\sqrt{2}\cdot\sqrt{A+C\mp\sqrt{A^2+B^2+C^2-2AC}}}{A+C\pm\sqrt{A^2+B^2+C^2-2AC}}$$

$$\phi = \text{Arctan}\left(\frac{C-A\mp\sqrt{A^2+B^2+C^2-2AC}}{B}\right)$$

$$\mathbf{A} = \frac{2\pi i}{\sqrt{B^2-4AC}}$$

$$\kappa_a = \frac{A+C\pm\sqrt{A^2+B^2+C^2-2AC}}{\sqrt{2}\cdot\sqrt{A+C\mp\sqrt{A^2+B^2+C^2-2AC}}}$$

$$\kappa_b = \frac{A+C\mp\sqrt{A^2+B^2+C^2-2AC}}{\sqrt{2}\cdot\sqrt{A+C\pm\sqrt{A^2+B^2+C^2-2AC}}}$$

Interactive Applet

Appendix Applet 6a on the CD-ROM sketches the conic determined by

$$Ax^2 + Bxy + Cy^2 = 1$$

and provides on-the-fly values of all the above formulas.

7. Conics As Paths of Celestial Bodies. From the perspective of the planets in our solar system, the sun's gravitational field closely approximates that of a fixed point of mass M. In the formulas, we assume M is at the origin $\mathbf{O}$. Let m be a point of mass much less than that of M, such as a planet. With m orbiting M, take at some arbitrary instant a reading of m's position vector $\mathbf{R} \neq \mathbf{O}$ and velocity vector $\mathbf{V}$. We assume that these vectors $\mathbf{R}$ and $\mathbf{V}$ lie in the (x, y)-plane but not on the same line. The strength of the force between M and m is given by Newton's law of gravitation

$$F = G\frac{Mm}{R^2},$$

where R is the length of $\mathbf{R}$. For convenience, we assume here, and in the formulas, that

$$\textbf{Units have been chosen so that the force strength is } F = \frac{1}{R^2}.$$

This central inverse-square force causes m to trace out a path lying on a conic determined by $\mathbf{R}$ and $\mathbf{V}$, and it turns out that every ellipse, parabola or hyperbola is determined (up to translation) by a suitable choice of initial position and velocity. Using polar coordinates, we may write $\mathbf{R} = (R, \alpha)$ and $\mathbf{V} = (V, \alpha + \beta)$. The setup is illustrated in Figure A1.4:

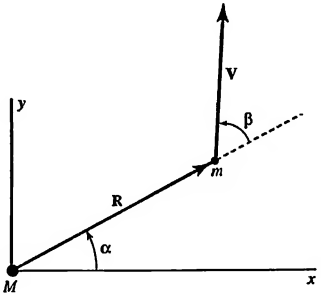


FIGURE A1.4. M is a fixed point-mass, and m is a moving point with a much smaller mass. This picture can be regarded as a snapshot of m in mid-flight, or equally well, as giving initial conditions or "launch data" for m.

Every conic having a focus at the origin is determined by some selection of four scalars R, α, V, β. Because the focus is fixed at (0, 0), these conics have only $5-2=3$ degrees of freedom, so there must be some slack within the family $\{R, \alpha, V, \beta\}$. Correspondingly, for any conic, there is within $\{R, \alpha, V, \beta\}$ a one-dimensional family of initial conditions, each member determining the same conic. It is obvious what this family is: as m moves along its one-dimensional path, we can at any instant read its position and velocity. We know from elementary physics that these readings all specify one and the same conic, and all readings belong to the family. In this sense, we say that $\{R, \alpha, V, \beta\}$ contains one dimension of redundancy or slack.

Equation (A1.2) [p. 363] tells us that any conic with a focus at the origin can be written in polar coordinates as

$$r = \frac{p}{1 - \varepsilon \cdot \cos(\theta - \phi)},$$

where the semi-latus-rectum p determines the conic's size, the eccentricity gives ε its shape, and ϕ specifies its tilt. All three of these parameters can be expressed in terms of R, α, V, β. All four of these numbers change as m moves in its orbit. Interestingly, there are also two other dynamical quantities—angular momentum h with respect to the origin, and total energy E—that are determined by R, α, V, β but, like the conic itself, *remain constant through time.* In fact, h alone determines the conic's size (latus-rectum), and E and h together define the conic's shape (eccentricity).

To make the following formulas involving energy more meaningful, let's briefly outline how we compute the various energies of a particle in M's force field. Let m initially be at rest at some location so far from M that only negligible work (force through distance) is needed to move m indefinitely further away from M. (From the standpoint of numerical calculations, we're taking m to be "at infinity.") By convention, we take the total energy of m at this resting point to be $E = 0$. Upon releasing m, it slowly moves toward M, gradually picking up speed, its total energy splitting up into positive and negative parts: $E = KE + PE$. Here KE is the energy of motion (kinetic energy) and PE is the potential

energy of position. The increasing speed gives m more energy of motion, and this has the positive value $KE = \frac{1}{2}mv^2$. Its decreasing distance from M means that symmetrically, m has negative potential energy: $PE = -\frac{1}{2}mv^2$. In this regard, m is similar to a brick dropped from a tall building—its position in the earth's gravitational field at the top of the building enables the brick to use potential energy to fuel its speed increase as it hurtles downward. As m falls toward M, it goes into ever greater "PE-debt" to feed this increase in KE.

If at some location P, m hits a perfectly-reflecting wall, its rebound speed is just enough to send it back to its original "infinite" distance from M. However, suppose you intercede, slowing m down at P This removes some of its kinetic energy, making the total energy negative; m then can't make it all the way back to infinity, and instead becomes trapped in an elliptical orbit around M. The more energy is removed, the smaller the orbit. On the other hand, if at P you *add* speed to m, you thereby endow it with a net positive total energy, giving m enough energy not only to escape from any orbit around M, but to travel "beyond the line at infinity." In this journey, m travels along a hyperbola branch.

The various measurements of conics are given as formulas in R, α, V, β—scalars defining position and velocity at any particular instant. When possible, we include formulas in terms of angular momentum h and total energy E, too. (Two standard references on this material are [14] and [15].)

$$a = \frac{R}{2 - RV^2} = \frac{-1}{2E}$$

$$b = V\sqrt{\frac{R^3 \sin^2\beta}{2 - RV^2}} = \frac{h}{\sqrt{-2E}}$$

$$\|F\| = R\frac{\sqrt{1 - RV^2(2 - RV^2)\sin^2\beta}}{2 - RV^2}$$

$$\|D\| = \frac{R}{(2 - RV^2)\sqrt{1 - RV^2(2 - RV^2)\sin^2\beta}}$$

$$\|F\| = \frac{\sqrt{2Eh^2 + 1}}{2E}$$

$$\|D\| = \frac{1}{2E\sqrt{2Eh^2 + 1}}$$

$$\varepsilon = \sqrt{1 - RV^2(2 - RV^2)\sin^2\beta} = \sqrt{2Eh^2 + 1}$$

$$p = R^2V^2\sin^2\beta = h^2$$

$$\phi = tilt = \alpha - \text{Arctan}\left[\frac{RV^2\sin(2\beta)}{2(RV^2\sin^2(\beta) - 1)}\right]$$

$$\mathbf{A} = \pi V\sqrt{\frac{R^5\sin^2\beta}{(2 - RV^2)^3}} = \frac{\pi h}{(-8E)^{3/2}}$$

$$\kappa_a = \frac{1}{R^2V^2\sin^2\beta} = \frac{1}{h^2}$$

$$\kappa_b = V^2(2 - RV^2)^2\sin^2\beta = h(-2E)^{3/2}$$

Note: $(h, k) = (\|F\| \cdot \cos\phi, \|F\| \cdot \sin\phi)$, which is quite compact. The above formulas for $\|F\|$ and ϕ of course allow us to express (h, k) in terms of R, α, V, β explicitly, but this expression is considerably longer.

Fixing any three of the parameters R, α, V, β and varying the fourth defines a family of conics. These conics can be plotted using, say, the polar equation in (A1.3), since we have formulas for p, ε and ϕ in terms of R, α, V, β. Figure A1.5 illustrates two possibilities.

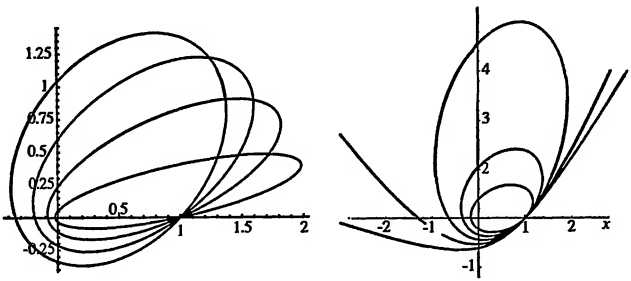


FIGURE A1.5. The left picture shows four different orbits corresponding to fixing all launch parameters except the angle β in Figure A1.4. We have taken $r = 1$, $V = 1$, $\alpha = 0$, and have let β assume the four different values

$$\frac{\pi}{16}, \frac{\pi}{8}, \frac{3\pi}{16}, \frac{\pi}{4}.$$

The right picture shows five conics determined by fixing $r = 1$, $\alpha = 0$, $\beta = \frac{\pi}{4}$ and successively launching the particle with speeds V of 0.75, 1.00, 1.25, 1.5 and 1.75. These increasing launch speeds result in increasing eccentricities and angles of tilt. They are, respectively: $\varepsilon \approx 0.28$, $\phi \approx 21.1°$ (smallest ellipse); $\varepsilon \approx 0.51$, $\phi \approx 46.15°$ (middle ellipse); $\varepsilon \approx 0.78$, $\phi \approx 74.48°$ (largest ellipse); $\varepsilon \approx 1.12$, $\phi \approx 99.8°$ (inner hyperbola branch); $\varepsilon \approx 1.53$, $\phi \approx 109.53°$ (outer hyperbola branch).

Interactive Applet

Appendix Applet 7a on the CD-ROM generalizes Figure A1.5 and provides on-the-fly values of all the above formulas.

8. Conics Through Five (or Three) Points. A set of five distinct points in the plane almost always uniquely determines a conic. The "almost always" arises from the possibility of degenerate conics. For example, five collinear points can lie on a double line, on two crossing lines, or on two parallel lines. In the following we assume that the five points lie on a nondegenerate conic. Each point $P_i = (a_i, b_i)$ on the conic of course satisfies the conic's quadratic equation, and from these five equations we can determine that equation. These five equations are linear in the unknown coefficients $A, \ldots, F$; we write

$$a_i^2 A + a_i b_i B + b_i^2 C + a_i D + b_i E + F = 0 (i = 1, \ldots, 5) \qquad \text{(A1.12)}$$

If we select coordinates so that none of the five points is the origin, we can take $F = 1$. One can solve the system (using Maple or Mathematica, for example). The coefficients $A, \ldots, E$, now being known, the conic can be plotted. If the points are selected to be pairwise symmetric about the origin, then there are only three freely-chosen points. In this case, $D = E = 0$, the conic's center is at the origin and the linear system involves only the unknowns A, B, C. If the symmetrically-chosen points are $\pm(a_1, a_2), \pm(b_1, b_2), \pm(c_1, c_2)$, then A, B, C are given by

$$A = \frac{a_1 a_2 (b_2^2 - c_2^2) + b_1 b_2 (c_2^2 - a_2^2) + c_1 c_2 (a_2^2 - b_2^2)}{(a_1 b_2 - b_1 a_2)(a_1 c_2 - c_1 a_2)(b_1 c_2 - b_2 c_1)},$$

$$B = \frac{(a_1^2 b_2^2 - a_2^2 b_1^2) + (c_1^2 a_2^2 - c_2^2 a_1^2) + (b_1^2 c_2^2 - b_2^2 c_1^2)}{(a_1 b_2 - b_1 a_2)(a_1 c_2 - c_1 a_2)(b_1 c_2 - b_2 c_1)},$$

$$C = \frac{a_1 a_2 (c_1^2 - b_1^2) + b_1 b_2 (a_1^2 - c_1^2) + c_1 c_2 (b_1^2 - a_1^2)}{(a_1 b_2 - b_1 a_2)(a_1 c_2 - c_1 a_2)(b_1 c_2 - b_2 c_1)}.$$

In either the three-point or five-point case, all the formulas in the quadratic-equation list can be applied to find area, eccentricity, and so on.

Animated Applets

Appendix Applets 8a and 9a on the CD-ROM sketch the conic determined by five points and give on-the-fly values of the formulas above.

Interactive Applet

For an animated, interactive version of Figure A1.6, see Appendix Applet 10a on the CD-ROM.

Given five points in the plane, it can be surprisingly difficult to guess at which conic goes through them. Consider, for example, Figure A1.6.

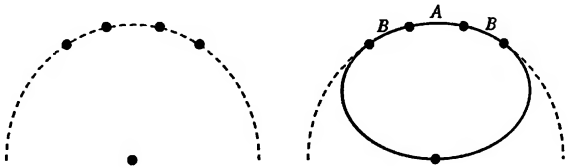


FIGURE A1.6. The top four points lie on the dashed circle; the fifth is the circle's center. Yet there is a conic passing through all five points. This stumps many people, until they see the "obvious" solution—the ellipse shown on the right. Interestingly, in the middle area labeled A, the ellipse falls below the circle by less than one-half of one percent of the circle's radius. It rises a similarly tiny amount above the circle in the two B regions. By Bézout's theorem, the circle and ellipse intersect in just these $2 \cdot 2 = 4$ points, and in no others.

Imaginary Curvature

Imaginaries can occur in the formulas for the curvature κ_b, and one can use the parametric equations $x = a\cos t$, $y = ib\sin t$ to derive curvature formulas for the ellipse in Figure 6.1 on p. 92 in the same way as with an ordinary ellipse. One continues to define curvature as the amount of turning per arc length, $\frac{d\theta}{ds}$; straightforward computation gives ds and $d\theta$. For ds,

$$ds = \sqrt{dx^2 + dy^2} = \sqrt{a^2\sin^2 t + (ib)^2\cos^2 t}\,dt.$$

And for $d\theta$,

$$\text{slope} = \frac{dy}{dx} = -\frac{ib}{a}\cot t.$$

If we write

$$\theta = \text{Arctan}\left(-\frac{ib}{a}\cot t\right),$$

then $d\theta$ becomes, after a little simplifying,

$$d\theta = \frac{iabdt}{a^2\sin^2 t + (ib)^2\cos^2 t}.$$

Therefore a general curvature formula for the ellipse in Figure 6.1 is

$$\kappa = \frac{d\theta}{ds} = \frac{iab}{(a^2\sin^2 t - b^2\cos^2 t)^{3/2}}. \tag{A1.13}$$

At $t = 0$, (A1.13) becomes

$$\kappa_0 = -\frac{a}{b^2};$$

and at $t = \frac{\pi}{2}$, (A1.12) gives

$$\kappa_b = \frac{ib}{a^2}.$$

Both of these agree with our formulas in the parametric equations list when b is replaced by ib there. In the other lists, formulas for the semi-axes occur and lead to the curvature formulas via these ratios.

Four "Grand Sweep" Pictures

In Chapter 8, we met a "grand sweep" picture—the family of conics in Figure 8.10 on p. 136, generated by holding size and orientation fixed and varying the shape parameter ε = eccentricity. Such a picture provides a nice overview of how the variously-shaped conics fit together. Since ε^2 runs through the reals in Figure 8.10, the picture shows conics corresponding to both real and imaginary eccentricity.

These grand sweep pictures are obtained from our formula lists by fixing the size parameter (orientation ϕ is assumed 0 throughout), and allowing the appropriate shape parameter to vary. These formula lists generate four distinct pictures. Here they are:

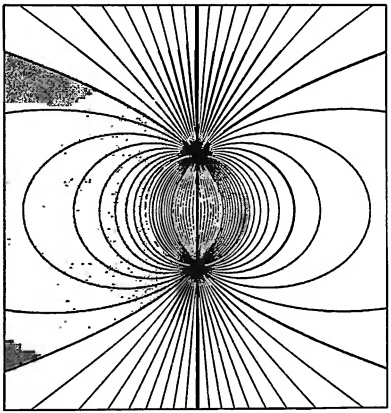


FIGURE A1.7. This is the family of projected cone-sections generated by varying the angle ψ of the slicing plane—the situation suggested by Figure A1.2. We take the size parameter H to be 1; starting from 0°, ψ is successively incremented or decreased by 5°. The parabolas bounding the shaded regions correspond to ±45°. We have seen that all these cone-sections are given equally well by the polar-coordinate formulas (A1.2) and (A1.3)—algebraic counterparts of Figure A1.2. In this case, the size parameter $H = 1$ is the semi-latus rectum $p = 1$. For the shape parameter we use not ψ, but rather the plane's slope $= m = \tan\psi$, which is equal to the eccentricity ε.

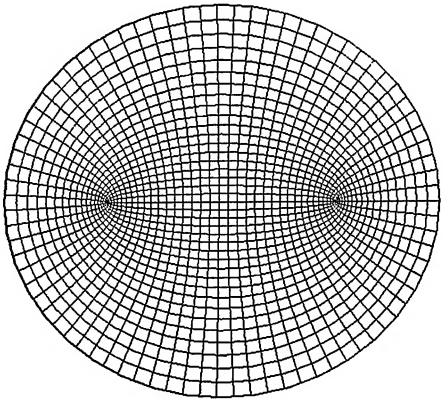


FIGURE A1.8. This "confocal conics" picture is quite famous. It appears in Figure 8.11 on p. 137 and also as a system of fringe-curves in Figure 17.2 on p. 343. In yet another guise, it is a grand sweep picture for stakes-and-string conics. The "eyes"—the foci—are the two stakes, fixed at $(\pm 1, 0)$. The string length is $2L$, and L parametrizes shape; our formulas show that it is

$$L = \frac{1}{\varepsilon}.$$

This fits in with what we see: the longer the string, the more nearly circular the ellipse. These ellipses may be written in the standard form

$$\frac{x^2}{L^2} + \frac{y^2}{L^2 - 1} = 1.$$

This is a natural invitation to let L^2 run through the reals. For "impossible" ellipses, when $L^2 < 1$, the standard form makes it easy to see that these ellipses are in fact hyperbolas.

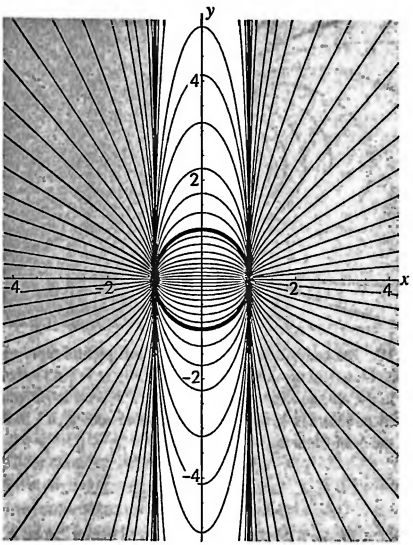


FIGURE A1.9. The picture in Figure 8.10 (p. 136) and Figure 11.10 (p. 198) rears its head yet again! Here we see it as the family of quadratic-equation conics in standard form:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1.$$

We treat a as size parameter; fixing $a = 1$, we can then use b as the shape parameter. We let b^2 run through all reals; the shaded area corresponds to negative b^2—that is, imaginary b. This picture also shows conics in the list for parametric equations $x = a\cos t$, $y = b\sin t$, since eliminating t returns us to

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1.$$

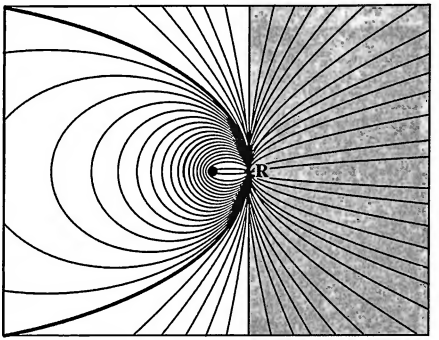


FIGURE A1.10. The heavy dot at the center of this picture represents a fixed "sun." A spaceship or projectile launched vertically at R with speed V traces out a conic in the unshaded portion of the picture. If we choose units of measurement so that the sun is 2 units away from the launching pad, then the polar equation of the orbit becomes

$$r = \frac{4V^2}{1 - \sqrt{1 - 4V^2(1 - V^2)} \cdot \cos\theta}.$$

The escape speed is $V = 1$, the escape trajectory thus lying on the heavily-shaded parabola opening to the left. The vertical distance from the sun to any conic on the left is the semi-latus-rectum—physically, the (constant) angular-momentum-squared of the projectile. This quantity is also the path's radius of curvature at the launching pad. The above polar equation allows us to sketch curves corresponding to *all* real values V^2. The curves in the shaded half correspond to negative V^2—"imaginary speed." Physically, these orbits—all hyperbolas—are traced out if the force is *repelling* rather than attracting. An example: let a fixed electron play the role of the sun, and an electron free to move play the role of spaceship. The force between them is inverse-square and repelling; the free electron will move along a hyperbola in the shaded-area family.

Appendix 2 Topology: A Quick Handshake

Introduction

This appendix supplies a practical understanding of the topological ideas used in this book. The kind of topology we use is mainly "rubber band geometry" and "rubber sheet geometry." Two objects are topologically equivalent if there's an invertible map from one object to another that is continuous in both directions—"bicontinuous." For example, one can continuously stretch a circle to an ellipse, or a sphere to an ellipsoid, and in each case one can play the scenario backwards, seeing the ellipse continuously deform back to a circle or the ellipsoid to a sphere. The circle and ellipse are *one-dimensional*, in that we can think of them as mathematical idealizations of objects made of thread or thin strings of rubber. The sphere and ellipsoid are *two-dimensional*—we can think of those as being constructed from ideal sheets of rubber, which we can stretch, bend or glue together to make various two-dimensional surfaces. (An isolated point has dimension zero, and a hunk of clay has dimension three.)

Topology is at one end of a rather broad spectrum of geometries, ranging is degrees of "strictness." For example, Euclidean geometry is quite strict in that its notion of equivalence is *congruence*—two objects are equivalent if and only if you can get from one to the other by using parallel translation and/or rotation and/or reflection. You might call it "stiff cardboard geometry." A more relaxed notion of equivalence is *similarity*, in which size doesn't matter. In "similarity geometry" we may pick up the piece of cardboard, possibly flip it over, and throw it back onto the plane where it will land translated and rotated, and additionally, the object may be uniformly magnified or contracted, as in zooming in or out on a computer monitor. This makes the original and transformed objects *similar*, as understood in Euclidean geometry Still less demanding is *perspective* or *projective* geometry, where we're allowed to take a snapshot of the cardboard from any angle, other than edge-on. (This is like dispensing with the viewing plane altogether, and viewing the cardboard objects in space!) As a geometry, topology is yet more "relaxed": we can turn the cardboard into rubber and allow it to deform within reasonable limits.

Historically, one of the first examples of a rubber-band question is the "Bridges of Königsberg Problem." The Pregel River divides Königsberg, in Prussia (now the city of Kaliningrad in Russia), into four separate land masses connected by seven bridges.

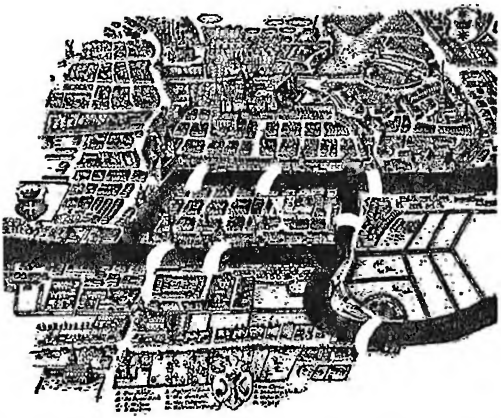


FIGURE A2.1. The Pregel River divides the town into a central "island" and three other regions to the north, south and east. For clarity, the map's seven bridges have been highlighted in white.

These bridges offered a pleasant way to take a stroll, and many townsfolk opted to cross all seven bridges, starting and ending at home. But it seemed that you could never make a round trip by crossing each bridge exactly once—you either had to omit a bridge, or you had to cross some bridge more than once. Was this because nobody had ever stumbled upon just the right route? This curious phenomenon finally reached the great Swiss mathematician Leonhard Euler, who of course attacked the headscratcher mathematically. He concluded that it was indeed impossible. How did he do it?

He began by idealizing the land masses to points, and the bridges to arcs, as suggested in Figure A2.2. He viewed the question as a problem in genuine "rubber band geometry." His solution was novel and represented a real conceptual breakthrough. Euler recognized that the problem went deeper than "distance geometry"—it wasn't the length of the bridges that mattered, but rather the way in which they were interconnected.

Later, Euler also noticed a "rubber geometry" phenomenon of one higher dimension. This, a rubber sheet theorem, can be stated in quite familiar terms: take a solid cube of wood, and repeatedly pass it through a band saw, cutting off as many little pieces as you like. At the end, you'll have a "convex polyhedron" with perhaps quite a great number of vertices V, edges E, and faces F. But no matter how big or small the numbers V, E or F are, in every instance they satisfy $V - E + F - 2 = 0$. This is a "rubber sheet fact" because you could just as well replace the wooden convex polyhedron by one made of a pliable material, and then squeeze or stretch it as much or as little as you like, always making sure the process is bicontinuous. For example, you could sketch the vertices, edges and faces on an ordinary sphere-shaped balloon, and the relation still holds. You can even stretch the pliable material out to make it look like a long party-balloon—but the two-way

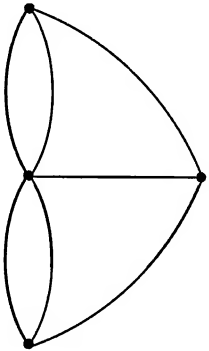


FIGURE A2.2. Euler's inspiration was to shrink the four land masses to four points (drawn heavily, here) and to idealize the seven bridges to seven arcs connecting them. The problem then reduced to this: without ever lifting up the pencil, can you draw a roundtrip path from any point to itself, traverseing each arc exactly once? He found that the answer is no.

continuity means you're not allowed to connect the balloon's ends to make a torus shape, for example. If you did that, essentially drawing vertices, edges and faces on an inner tube, you'd find that the relation changes to $V - E + F = 0$. Draw vertices, edges and faces on a surface with g holes in it—a surface of genus g—and you find that the general relation is $V - E + F - 2 = 2g$. None of V, E, F, or g are changed under any bicontinuous transformation—the theorem is genuinely one in topology.

Some Definitions

In observing the physical world at everyday sizes, it seems that almost everything changes in a continuous way. Even when things happen too fast for the human eye to follow—like a bullet plowing through a cinder block—high speed photography shows us in convincing slow motion that the change is in fact continuous. Yet historically, capturing the essence of continuous processes eluded thinkers from the ancient Greeks on. It therefore represented a real victory when Augustin-Louis Cauchy (1789–1857) devised a definition of continuity that is both simple and rigorous. It has withstood the test of time, and is actually the germinal notion of topology. Think, for a moment, of Cauchy's standard definition of continuity for a function $y = f(x)$ from $\mathbb{R}$ to $\mathbb{R}$. We say "f is continuous at a point $a \in \mathbb{R}$ if, for any arbitrarily small positive ϵ, there's a positive δ such that $|x - a| < \delta \Rightarrow |y - f(a)| < \epsilon$." Very importantly, $|x-a| < \delta$ and $|y-f(a)| < \epsilon$ each define an open interval in $\mathbb{R}$—that is, a line segment without endpoints. Cauchy's definition actually comes down to a little "contest" involving open intervals—my challenger chooses a very small open interval, and if f is continuous, I can always counter by producing an open interval mapping under f *into*

my friend's open interval. The objects used to define continuity of functions $f: \mathbb{R} \to \mathbb{R}$ are open intervals in $\mathbb{R}$.

In "rubber geometry," transformations are assumed to be bicontinuous, and that means we need a definition of continuous function that is more general than only from $\mathbb{R}$ to $\mathbb{R}$. As a first step in this direction, we generalize from open interval to open *set*. The open intervals form a "basis" of all open sets of $\mathbb{R}$. That is to say, an **open set of** $\mathbb{R}$ is any union of (possibly infinitely many) open intervals. This, of course, implies that an arbitrary union of open sets is open. One can also show that the intersection of any finite number of open sets is open. This includes the empty union and empty intersection, so the empty set and the whole line $\mathbb{R}$ are also open.

These ideas extend in a natural way to $\mathbb{R}^n$. Just as an open interval of $\mathbb{R}$ consists of all points strictly inside a "zero-dimensional circle" such as $(x-a)^2 = \delta$ (a zero-circle thus consists of two points), in $\mathbb{R}^2$, a typical basic open set can be taken to consist of all points strictly inside a 1-circle, which is an open disk. In $\mathbb{R}^n$, the basic open sets similarly consist of open n-balls. **Open sets of** $\mathbb{R}^n$ are defined as arbitrary unions of basic open sets. In all cases, we say *the open sets form a topology on the space.* In 1914, Felix Hausdorff abstracted these ideas into a simple, general axiomatic form which can be stated this way:

A topological space is a set T together with a collection of subsets closed under arbitrary unions and finite intersections. (This collection therefore includes T and the empty set $\emptyset$.) The subsets are called open sets.

This definition is very flexible. For example, if S is a subset of T, then S *inherits* a topology from T: its open sets are merely the intersections of S with T's open sets, and it can easily be shown that this in fact does define a topology on S.

A set N in a topological space T is a **neighborhood of a point** $P \in T$ provided N contains an open set of T with P in that open set. If we wish to make the reference to T explicit, we may call N a T**-neighborhood of** P.

One can now define continuity for a function from one topological space T_1 to another topological space T_2:

A function $\mathbf{F} : T_1 \to T_2$ **is continuous at a point** $P \in T_1$ if this holds: *for any neighborhood N_2 of the point $\mathbf{F}(P)$ in T_2, there exists a neighborhood N_1 of P in T_1 such that $\mathbf{F}(N_1) \subseteq N_2$.* **F** *is* **continuous on** T_1 *if it is continuous at each point of* T_1.

We can now give a precise definition of topological "sameness":

Two topological spaces T_1 and T_2 are **equivalent** *(or* **homeomorphic***) if there exists a* 1:1*-onto mapping $\mathbf{F} : T_1 \to T_2$ such that* **F** *is continuous on T_1, and its inverse $\mathbf{F}^{-1}$ is continuous on T_2.*

This definition implies that the open sets in T_1 match up exactly with the open sets of T_2 —that is to say, there's a 1:1 onto correspondence between the open sets which preserves arbitrary union and finite intersection.

We let "$\cong$" denote topological equivalence.

Five mathematicians who figured in the development of topology. Left to right, top to bottom:

Leonhard Euler (1707–1783);

Augustin-Louis Cauchy (1789–1857);

Georg Friedrich Bernhard Riemann (1826–1866);

Henri Poincaré (1854–1912);

Felix Hausdorff (1868–1942).

A point P in a topological space T is a **boundary point of a subset** $S \subseteq T$ *if every neighborhood of P contains points in S and points not in S. The union of S and all its boundary points is called* **the topological closure of** S **in** T; *if the meaning is clear from context, we sometimes call this* **the closure of** S **in** T, *or even simply* **the closure of** S.

Examples of topological equivalence:

- $\mathbb{R} \cong (0, 1)$. Let $\mathbf{F} : x \to 1 + \frac{2}{\pi}\text{Arctan}(x)$.
- $(0, 1) \cong (0, 2)$; $[0, 1] \cong [0, 2]$. In each case, let $\mathbf{F} : x \to 2x$.
- $\mathbb{R}^n \cong$ the open ball of points $(x_1, \ldots, x_n)$ in $\mathbb{R}^n$ satisfying $x_1^2 + \cdots + x_n^2 < 1$. Let

$$\mathbf{F} : (x_1, \ldots, x_n) \to \frac{1}{\sqrt{x_1^2 + \cdots + x_n^2 + 1}(x_1, \ldots, x_n)}.$$

The boundary of this open ball in $\mathbb{R}^n$ is the sphere S^{n-1}, the locus of $x_1^2 + \cdots + x_n^2 = 1$ in $\mathbb{R}^n$.

It can be harder to prove that two given spaces are *not* topologically equivalent. For our purposes, there's a quick way to gain an appreciation of the substance of topological equivalence: look at various examples that are equivalent (as in Figure A2.3) and others that are *not* (Figure A2.4).

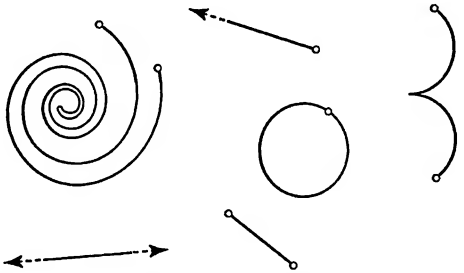


FIGURE A2.3. All the above topological spaces are homeomorphic. An open circle indicates a point not in the set.

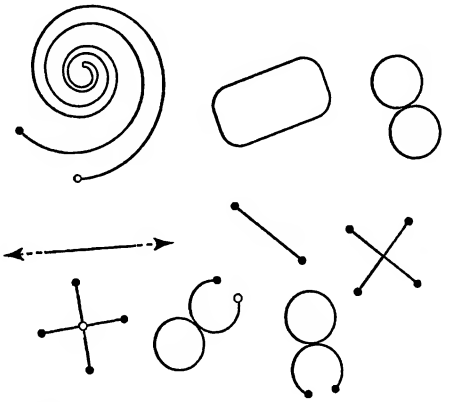


FIGURE A2.4. No two of the above topological spaces are homeomorphic. A filled-in circle indicates a point included in the set. All the topological spaces in this figure and in Figure A2.3 are one-dimensional.

Figure 1.7 on p. 10 provides some familiar examples of inequivalent topological spaces of two dimensions: the only equivalent ones there are the first two, which are both equivalent to a closed disk.

New Spaces from Old Ones

As happens so often in science, one can combine elementary "building blocks" to make more complex objects, a bit reminiscent of the way chemical elements bond in various ways to make compounds. In topology, these building blocks can be especially simple. Using finitely many single points, intervals or disks, one can construct a dazzling array of topological spaces. The following five examples provide some of the flavor. For our purposes, "construction" will mean simply "gluing" or "pasting," though other techniques are also routinely used by topologists.

Example. Consider the two intervals [0, 1] and [2, 3] in $\mathbb{R}$. Each interval inherits a topology from $\mathbb{R}$; the interval's open sets are its intersections with open sets of $\mathbb{R}$. Somewhat as with two pieces of wood, we can slide the two intervals until they touch, and glue them together at their common endpoint. We can do this with a so-called *projection map* π: this goes from the "before" pieces—that is, $[0, 1] \cup [2, 3]$, to the "after gluing" piece—that is, to [0, 2]. In our example, we let π be the identity map on the points of [0, 1], and translation one unit to the left on the points of [2, 3]. This map is obviously 1:1 on $[0, 1) \cup (2, 3]$. Notice that the right end of [0, 1] (the point $1 \in \mathbb{R}$) and the left end of [2, 3] (the point $2 \in \mathbb{R}$) both map to the final "glue seam", which is $1 \in \mathbb{R}$. We define a set O of [0, 2] to be *open* if and only if $\pi^{-1}(O)$ is open. This gives our "glued together" set the same topology as the usual interval [0, 2] has, and we say that *the two original points* 1 *and* 2 *have been identified to the final point* 1.

Example. From the standpoint of gluing together building blocks, the situation in topology is somewhat different from, say, atoms or Tinker toys: the rubbery nature of topological construction units means that points of possible attachment are *everywhere*. One can, for example, identify different parts of one and the same building block—not something atoms can do. A familiar example is a single closed interval: we can identify its endpoints by bringing them around to a position of coincidence, and define open sets around the fused point using an analogue of the projection map used a moment ago. In that example we simply obtained another closed interval, but now we get something topologically different—a topological loop.

Example. The idea of identifying points and defining open sets containing them applies equally well to the identifications used in Figure 1.7 on p. 10. In each case we bring points into coincidence so that single or double arrowheads match up, and then define open sets about identified points using a projection map. Although one can certainly identify opposite edges of a square to make a cylinder, to furthermore physically paste the remaining edges to get, say, a torus necessitates stretching the rubber sheet. This problem comes from working in the cramped quarters of $\mathbb{R}^3$ The problem becomes even more evident when one attempts to construct a Klein bottle or real projective plane, neither of which can be realized in $\mathbb{R}^3$ without both stretching and self-intersection. The Klein bottle *can* be constructed in $\mathbb{R}^4$ without self-intersection. The fact that $\mathbb{R}^3$ is so small is one reason for using identifications as in Figure 1.7: not only do we succeed in making new spaces, but instead of trying to imagine objects in $\mathbb{R}^4$, we merely need to think in terms of identifying points and creating

appropriate open sets containing them. In any case, we have to make some concessions because of our limited abilities to visualize.

Example. All identifications in Figure 1.7 are 2:1. One can easily imagine a 3:1 or an n:1 identification. One could in fact identify *a whole circle of points* to a single point. Example: squeeze the equator of a sphere so that all equatorial points draw together. In the limit, the circle degenerates to a single point, and the sphere becomes two spheres touching at that one point. We see this in Figure 12.11 on p. 231, which shows topologically what happens when the radius of a circle decreases to zero. We can play the same game with the cylinder in Figure 1.7. It has a circle at its top and bottom. Squeezing one of these circles to a point changes the cylinder to a closed disk. If we then squeeze the other circle to a point, the disk becomes a topological sphere.

Example. An interesting instance of the last example arises in arriving at the topology of a complex circle in Chapter 12. Equation (12.9) on p. 218 tells us how a hyperbola branch attaches to each point of a circle; this is illustrated in Figure 12.1, p. 219. Each branch is of course homeomorphic to an open interval, so as we swing around the circle, these branches sweep out a cylinder with top and bottom circles missing. From further considerations there, we know that all the branches approach the same two points. By adding in the top and bottom circles and squeezing each of them to a point, we arrive at a sphere, illustrated in Figure 12.2.

We now come to an important definition:

A subset S of $\mathbb{R}^n$ is a **topological m-manifold** *or a* **topological manifold of dimension** *m if each point $P \in S$ is contained in an open set of S homeomorphic to an open m-ball. (An open 1-ball is an open interval, an open 2-ball is an open disk; in general, an open m-ball is any set homeomorphic to the set of $(x_1, \ldots, x_m) \in \mathbb{R}^m$ satisfying $x_1^2 + \cdots + x_m^2 < 1$.)*

Examples. Topologically, an m-manifold looks like an open m-ball near each of its points. Thus $\mathbb{R}$, a circle, ellipse, or square are all topological 1-manifolds. On the other hand $\mathbb{R}^2$, a sphere, the surface of a cube and a torus are all topological 2-manifolds. But the union of two crossing lines is *not* a manifold, because all open sets around the point of intersection still contain little crosses. Similarly, the union of two intersecting planes is not a 2-manifold: open sets around intersection points contain parts of both planes. There are other examples of non-manifolds: the "alpha curve" shown in Figure 11.3 on p. 191; two touching circles; two touching spheres. Although the cusp curve shown in Figure 11.4 has a "kink," it is nonetheless a topological 1-manifold, though it's not "smooth" at the cusp. Note: It turns out that all the examples in Figure 1.7 on page 10 can be realized as topological 2-manifolds in some $\mathbb{R}^n$ (from the Whitney embedding theorem below); we will often refer to these simply as "topological 2-spaces."

It is easy to see that a circle, an ellipse and a square are topologically equivalent. In a topological context, *any* curve topologically equivalent to a circle is often referred to as a **topological loop**, or simply a **loop**.

Smooth manifolds. In our study of Gaussian curvature, derivatives appear in the formulas. This is an example where more than just a topological manifold is needed. For example,

a coffee cup with cylindrical sides and a bottom having the shape of the popular fractal known as a "Koch snowflake" is a topological manifold, but at no point on the cup's lateral side can one carry out the differentiations needed to find Gaussian curvature. (Interestingly, the Koch snowflake has finite area and infinite perimeter; it's sometimes joked that you can fill the cup with coffee, but you can't paint it.) Now the graph of a single-variable differentiable function intuitively looks smooth, and we key off this idea in defining "*smooth manifold*." (The assumptions on such a manifold are more than enough to permit computing Gaussian curvature.)

We begin with the notion of *smooth mapping*. Let U be an open set of $\mathbb{R}^n$, U' an open set of $\mathbb{R}^l$, and $\mathbf{F}$ a map from U to U'. Specifically, let's write $\mathbf{F} = (F_1, \ldots, F_l)$, where each $F_i(x_1, \ldots, x_n)$ is defined on U. We say that $\mathbf{F}$ is **smooth on** U if all partial derivatives of each F_i exist and are continuous throughout U. (This is quite a strong condition, since "all" means arbitrarily high orders of pure or mixed partials with respect to all of the x_j.)

This idea easily extends to arbitrary sets $S \subseteq \mathbb{R}^n$ and $T \subseteq \mathbb{R}^l$: $\mathbf{F}$ is **smooth on** S if for each point $x = (x_1, \ldots, x_n)$ in S, there is an open set U of $\mathbb{R}^n$ and a smooth map on U that agrees with $\mathbf{F}$ on S.

This allows us to define a "smooth equivalence" between S and T: we say S and T are **diffeomorphic** if there exists a one-to-one map $\mathbf{F}$ from S onto T such that $\mathbf{F}$ and its inverse $\mathbf{F}^{-1}$ are smooth.

A set $M \subseteq \mathbb{R}^n$ is a **smooth manifold of dimension** m if each $x = (x_1, \ldots, x_n) \in M$ has a neighborhood $W \cap M$ (W open in $\mathbb{R}^n$) that is diffeomorphic to an open m-ball. When m is 1, we call the manifold a **smooth curve**; when m is 2, we call the manifold a **smooth surface.**

For a smooth manifold M, any particular diffeomorphism $\mathbf{F}$ from an above open set $W \cap M$ of M onto an open m-ball is called a **coordinate patch** on M. Such a diffeomorphism smoothly and faithfully maps $W \cap M$ to an open set U in $\mathbb{R}^m$. Of course U inherits coordinates from $\mathbb{R}^m$, and $\mathbf{F}^{-1}$ then transfers those coordinates to $W \cap M$, thus endowing that open set with "local coordinates." Various of these coordinatized open sets in M may overlap—symbolically, $\mathbf{F}^{-1}(U) \cap \mathbf{G}^{-1}(V) \neq \emptyset$ for coordinate patches $\mathbf{F}$ and $\mathbf{G}$. $\mathbf{F}$ and $\mathbf{G}$

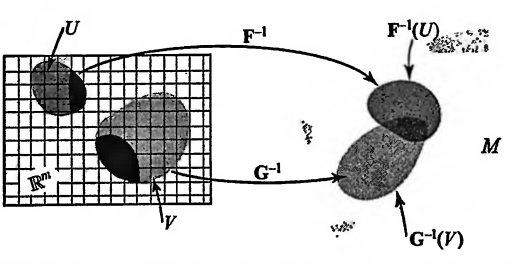


FIGURE A2.5. Diffeomorphisms locally transfer coordinates in $\mathbb{R}^m$ to open sets of the smooth manifold M. The images in M of these diffeomorphisms may overlap, and within the overlap, the image-coordinates usually won't agree. However, they are always smoothly related via diffeomorphisms. In this picture, $\mathbf{G} \circ \mathbf{F}^{-1}$ defines one such smooth map.

are smooth, and it can be verified that the composition $\mathbf{G} \circ \mathbf{F}^{-1}$ is smooth wherever it is defined. We can think of $\mathbf{G} \circ \mathbf{F}^{-1}$ as defining a "smooth gluing together" of the coordinate patches $\mathbf{F}$ and $\mathbf{G}$.

A diffeomorphism such as $\mathbf{F}^{-1}$ carrying a region U of $\mathbb{R}^m$ into M is called a **smooth parametrization of $\mathbf{F}^{-1}(U)$.**

Practical considerations. It is fairly straightforward to use the above definition of smooth manifold to show that any nondegenerate conic is a smooth manifold: in the real setting, it's a smooth curve in $\mathbb{R}^2$, and in the complex setting, it is a smooth surface in $\mathbb{C}^2 = \mathbb{R}^4$. It is also easy to see that a degenerate conic such as two crossing lines is not a smooth manifold, since the part around the intersection point isn't homeomorphic to an open set in $\mathbb{R}^m$. We have also considered curves such as the "alpha curve" and cusp curves in Chapter 11, and these, too, are not smooth manifolds. But to prove nonsmoothness directly from the definition, one must show that there exists *no* set of diffeomorphisms meeting our definition's requirements; in general terms, such a task can present difficulties.

Fortunately, all these examples are (plane) *algebraic curves* (introduced at the beginning of Chapter 11), and it turns out that for them, there are at most finitely many "bad" points; if these are removed, then in both the real and complex settings, the resulting set is a smooth manifold. Informally, one can call the nonbad points of the algebraic curve "good." The technical term for "good" is *nonsingular*; for "bad," it is *singular*. In the case of plane algebraic curves, there is a simple characterization of nonsingularity, and this characterization generally makes it easy to establish the nonsingularity of a particular point P of an algebraic curve. Here is the characterization: suppose the curve is defined by an irreducible polynomial p; then *in the locus $p(x, y) = 0$, a point P is nonsingular if and only if there is a line intersecting the curve at P with multiplicity* 1. We can apply this characterization to the origin of the cusp curve $y^2 - x^3 = 0$; we employ the algebraic method of finding intersection multiplicity outlined in Chapter 11: parametrize any line through the origin by $x = at$, $y = bt$ for some a, b. Substituting this into $y^2 - x^3$ gives $b^2t^2 - a^3t^3$, which has order at least 2 for all a, b; the origin of this cusp curve is therefore singular.

A basic generalization. The above definition of smooth mapping requires the existence and continuity of *all* partials of each F_i throughout some open set U of $\mathbb{R}^n$. One can weaken this assumption by requiring only that partials of each F_i exist *only up to some fixed order* k, and that these are continuous throughout U. (Any such partial thus has the form

$$\frac{\partial^m F_i}{\partial^{\alpha_1} x_1 \cdots \partial^{\alpha_n} x_n},$$

with $m \leq k$ and $\alpha_1 + \cdots + \alpha_n = m$.) One can then appropriately rewrite the above definitions, arriving at a differentiable manifold that one can call **smooth of order k,** parametrizations **smooth of order k,** and so on.

Smooth, smooth of order k, and this book. The minimum assumption on manifolds needed in this book is $k = 2$. However, nondegenerate conics and (nondegenerate) second-degree surfaces are all smooth, and our definition of smooth manifold fits in very naturally with existing literature and usage. Although there are many nonsmooth curves and surfaces that are smooth of order k, they are often somewhat artificial, such as the graph of the function

$$y = x^3 (x \geq 0),$$
$$y = 0 (x < 0).$$

Throughout the book's chapters, the reader may, if desired, safely replace "smooth manifold" by "manifold smooth of order 2."

Abstract smooth manifolds. "Smooth gluing" raises a basic question: instead of assuming from the start that the smooth manifold lies in some $\mathbb{R}^n$ (which is what we have done), couldn't one more generally define an "abstract smooth manifold" using coordinate patches that are smoothly glued together wherever they overlap? The answer is yes, provided the abstract space satisfies two Euclidean-like assumptions that essentially eliminate pathological spaces. The two assumptions are: there is a countable basis of the space; and the space is Hausdorff—that is, for any two points $P \neq Q$, there exist open sets U_P and U_Q with $U_P \cap U_Q = \emptyset$. This positive answer leads to another basic question: *are such abstract smooth manifolds actually more general than ones lying in an* $\mathbb{R}^n$? Remarkably, the answer is *no*. It was proved by Hassler Whitney in the *Whitney embedding theorem* that such an abstract manifold of dimension m can be smoothly embedded in the Euclidean space $\mathbb{R}^{2m}$. This sheds some light on the pictures in Figure 1.7 (p. 10); all those examples have dimension 2, but only the closed disk actually lives in $\mathbb{R}^2$. The cylinder, Möbius strip, sphere and torus must be viewed in $\mathbb{R}^3$ to see them as a smooth manifold in a Euclidean space. We are not that lucky with the real projective plane or the Klein bottle—neither of these can be embedded in $\mathbb{R}^3$ without introducing some non-smooth aspect, such as self-intersections. Whitney's $2m$-embedding theorem tells us, however, that the real projective plane *can* be realized as a smooth surface sitting in $\mathbb{R}^{2\cdot 2} = \mathbb{R}^4$. The same is true of the Klein bottle.

In a sense, Whitney's embedding theorem is the "grandfather of all embedding theorems." Through the years, mathematicians have found many others. They range from embedding various kinds of geometric spaces, to embedding algebraic structures into larger ones.

Oriented smooth manifolds. Another important notion is an *oriented manifold*, and for our purposes, we assume that the manifold is smooth. One very simple example is $\mathbb{R}^n$ itself. In fact, when $n = 1$—the real number line—the positive orientation is determined by which side of the origin contains $+1$. The "arrow" from 0 to 1 then specifies the positive direction. Now the singleton $\{1\}$ forms an ordered basis of $\mathbb{R}^1$, and 1 is the determinant of the 1×1 matrix of ordered basis vectors; from this especially trivial case we generalize, bit by bit, to a definition of oriented smooth manifold. Going up one dimension, we see that the rows of the 2×2 identity matrix constitute an ordered basis of $\mathbb{R}^2$. The determinant of this matrix is positive, and we correspondingly say that this ordered basis defines a *positive orientation on* $\mathbb{R}^2$. Switching the order of the basis gives

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

and multiplies the determinant by -1. This negative determinant defines a *negative orientation on* $\mathbb{R}^2$. Similarly, $\mathbb{R}^3$—and more generally $\mathbb{R}^n$—is given a positive orientation by an

ordered basis if the determinant of that ordered basis is positive; the space is assigned a negative orientation if the determinant of the ordered basis is negative.

The determinant of a nonsingular (real) matrix can be thought of as the signed volume of the parallelepiped defined by its row vectors. Consider $\mathbb{R}^3$, for example, with positive orientation defined by the 3×3 identity matrix. Take any one of the three row vectors and rotate it toward another row vector. As we do this, the area of the Base in Base $\times$ Height decreases, so the positive volume decreases. When the rotating vector coincides with the second one, the base area is zero, making the volume zero. Continued rotation makes the determinant negative, and therefore the orientation becomes negative. Alternatively, we could rotate any one of the vectors toward the plane of the other two, decreasing the Height in Base $\times$ Height. Again, the positive volume decreases, becoming zero when the rotating vector lies in the plane of the other two. Continued rotation gives a negative determinant and reverses the orientation of $\mathbb{R}^3$.

Determinants also allow one to compare the orientations of two oriented $\mathbb{R}^n$'s. If the rows of a nonsingular matrix A define an orientation of a real vector space V_A^n, and B defines an orientation of a V_B^n, then the nonsingular matrix BA^{-1} naturally defines a map from V_A's basis to V_B's. Now $\det(B) = \det((BA^{-1})A) = \det(BA^{-1}) \cdot \det(A)$, so $\det(A)$ and $\det(B)$ *have the same sign precisely when* $\det(BA^{-1})$ *is positive.* This comparison tool is important, because in a smooth manifold consisting of coordinate patches smoothly glued together, each of those patches can be given an orientation. To define an orientation on the entire manifold, we need to compare orientations of overlapping patches: if the manifold is orientable, these local orientations must agree. Intuitively, one can look at overlapping coordinate patches as composed of extremely small-faced polytopes, each face being a small part of the tangent space to the manifold there. The matrices mapping between two bases arising from the coordinates of two overlapping patches vary from point to point, and are given by the Jacobian matrix

$$J(x_1, \ldots, x_m) = \left(\frac{\partial H_i}{\partial x_j}\right),$$

where $\mathbf{G} \circ \mathbf{F}^{-1}$ in Figure A2.5 is denoted by $\mathbf{H} = (H_1, \ldots, H_m)$. (Think of J as a varying version of BA^{-1}.) If $\det(J)$ is positive throughout the entire neighborhood, then the orientations assigned to each tangent space (intuitively extending a small polytope face) agree at each point of that neighborhood. If this condition holds throughout the overlapping part of every pair of coordinate patches making up the smooth manifold, then we say **the smooth manifold is orientable.**

$\mathbb{R}^3$ or $\mathbb{R}^4$? The notion of orientability allows us to say more about which surfaces can smoothly fit in $\mathbb{R}^3$, versus those which require $\mathbb{R}^4$ If the surface is orientable, such as a sphere, torus or cylinder, it can be realized in $\mathbb{R}^3$ without self-intersections or creases. The same is true if the surface has a boundary, examples being a Klein bottle with a disk removed, or the real projective plane with a disk removed. (This last is topologically a Möbius strip.) But if the surface is closed in the sense that it has no boundary, and if it is not orientable, then you *cannot* smoothly embed it in $\mathbb{R}^3$. Whitney's theorem guarantees, however, that $\mathbb{R}^4$ *does* suffice—that is, smooth surfaces never require $\mathbb{R}^5$ or larger. The full Klein bottle or real projective plane are two examples where $\mathbb{R}^4$ is necessary, but there

are infinitely many others. For example, take any sphere with g handles, draw an oriented circle around any one handle, and cut to separate. Then reattach the cut ends by stitching along the circles in mutually opposite senses; this destroys orientability. By removing a small disk from the surface, thereby creating a boundary, you can physically carry out this stitching in $\mathbb{R}^3$ by snaking one cut end through the hole. But if there's no such hole, the surface is closed, and $\mathbb{R}^4$ is required.

Suggestions for Further Reading

In the following suggestions for further reading, we make no attempt at completeness. Some of them relate especially to certain chapters of this book—for example references [9], [12], [13] for the curvature chapters, and [5], [8], [14], [15] for Chapter 16.

1. Apollonius of Perga, *Conics, Books I–III*, Green Lion Press, Sante Fe, NM (1997).

   This is a translation of Apollonius' epic work. The first three books of eight are included here, and its over 150 propositions and proofs, usually accompanied with a diagram or two, give a very good flavor of what the great man did with the paucity of tools available to him.

2. J. Coolidge, *A History of the Conic Sections and Quadric Surfaces*, Dover reprint (1968) of 1945 original.

   This is an indispensable resource for the history of conics. The book is scholarly but easy to read, and gives a good sense of what some of the geometrical giants of the past actually accomplished. An invaluable, honestly-written reference.

3. J. Downs, *Practical Conic Sections*, Dover reprint, 1993.

   Physically, conic sections arise in both nature and on the drafting table. This 100-page book gives different physical approaches to how they can be created: nine for ellipses, eight for hyperbolas and 10 for parabolas. The author includes optical/acoustical applications, Kepler's laws (including orbital velocities), and an entire chapter on what he calls "the directing circle." This is related to both the directrix and the "velocity diagram" appearing in Feynman's *Lost Lecture*. This highly-geometric book is well worth owning.

4. H. Eves, *An Introduction to the History of Mathematics*, 5th Ed., Saunders, New York, NY, 1983.

   This deservedly famous history is a well-balanced, highly-readable account from primitive counting to more recent, abstract mathematics. Four of its fifteen chapters concern ancient Greek mathematics. Throughout the book, exercises drawn from original problems of the day afford great hands-on experience and help make the history of mathematics come alive.

5. D. Goodstein and J. Goodstein, *Feynman's Lost Lecture: The Motion of Planets Around the Sun*, W. W. Norton and Company, New York, NY, 1996.

This is a brilliantly clear retelling of a 1964 lecture given by Feynman on planetary motion. Using 30-year-old notes, audio tapes and photos, the authors have reconstructed the "Lost Lecture," allowing one to experience the exciting originality of Feynman himself. The arguments are elementary and can be followed by almost anyone with a little patience. A winner.

6. K. Kendig, *Elementary Algebraic Geometry*, Springer-Verlag, New York, NY, 1977.

   This concrete introduction was written to help "put geometry back into algebraic geometry." In it, you will see in greater detail the intuitive ideas introduced in the present book about multiplicity and Bézout's theorem. It also includes a proof of the genus formula. A friendly introduction to algebraic geometry.

7. K. Kendig, Is a 2000-Year-Old Formula Still Keeping Some Secrets?, *Amer. Math. Monthly* **107** (2000) 402–415.

   Although this paper is about Heron's formula for the area of a triangle, the ideas in it were the germ for this book, *Conics*.

8. J. Marion and S. Thornton, *Classical Dynamics of Particles and Systems*, Harcourt Brace Jovanovich, New York, NY, 1988.

   A detailed, clearly-written graduate text that can also be used as a great resource for finer points that other texts often omit. For our purposes, the treatment of central force motion leading to conical paths is very germane. There are also enlightening discussions about apsidal angles, precession, and stability of circular orbits.

9. J. Oprea, *Differential Geometry and its Applications*, Prentice Hall, Upper Saddle River, NJ, 1997.

   An introduction to differential geometry by an outstanding expositor. Suitable for students from about the senior level on, it includes Maple programs as an integral part of the book. These programs draw, compute, and help concretize abstract concepts—for example, you'll be able to draw geodesics on a torus, cone, paraboloid, hyperboloid, catenoid, the Poincaré plane, and others. It's easy to get hooked on differential geometry using this book. Includes a generous selection of exercises.

10. G. Salmon, *Treatise on Conic Sections*, Cheisea, New York, NY.

    A lovely old book written in an earlier generation. Salmon, 29 when he wrote this in 1848, carefully leads the reader through lines, circles, and then on to conics proper. There are a great number of pretty, classical results—you can almost see Salmon, perhaps together with his friends Cayley and Sylvester, deriving joy from them. Well worth browsing though this historic account.

11. D. Scher, *Exploring Conic Sections with The Geometer's Sketchpad*, Key Curriculum Press, Emeryville, California, 2002.

    This little gem was inspired by another gem—a book by one Frans van Schooten (penned some 350 years ago) which details, in drawing after drawing, a wide range of mechanical devices for creating conic sections. A number of these have been translated into computer constructions using The Geometer's Sketchpad. Scher's 87-page booklet also comes with a CD-ROM. The ease, generality and power some of the constructions help to make this a great learning tool.

12. M. Spivak, *A Comprehensive Introduction to Differential Geometry*, Publish or Perish, 1999.

Whole generations of differential geometers have learned from Spivak's clear, direct expositions. The original two-volume set written in 1970 has evolved and grown to five volumes, but they all contain concrete examples and enlightening explanations that help lay a solid foundation in differential geometry. One often sees the same idea looked at from various viewpoints—a real plus. The later editions also contain many helpful exercises.

13. D. Struik, *Lectures on Classical Differential Geometry*, Dover reprint of 2nd Ed. (1961), Addison Wesley Publ. Co., Inc., Reading, MA, 1988.

This classic is a highly-geometric, concrete introduction to differential geometry. Carefully and tightly written, its roots go back to the masters of a century ago, and as a nontrivial "first course," gives a solid foundation for continuing on to more general treatments.

14. J. Synge and B. Griffith, *Principles of Mechanics*, 3rd ed., McGraw-Hill, New York, NY, 1959.

This standby has withstood the test of time. It's a clearly-written, friendly account of traditional Newtonian mechanics. It's both a good textbook and a handy reference.

15. E. Whittaker, *A Treatise on the Analytical Dynamics of Particles and Rigid Bodies*, 4th Ed., Cambridge University Press, Cambridge, 1961.

This is probably the most famous account of classical dynamics ever published. Making its original appearance in 1904, it is scholarly, comprehensive, elegant—and slow reading. It repays careful study.

Applets

System Requirements

PC

Speed: 1 GH or faster (Pentium 4, Athlon XP processor, for example).

Operating System: Windows NT/2000 or Windows XP or later.

Memory: At least 128 MB RAM; 256 or more is recommended. Memory/Bus speed of at least 333 MH. (Automatic with Pentium 4 or Athlon XP.)

Browser: Internet Explorer 6 SP1 or later. Netscape is *not* recommended for running these applets. **Note:** For security reasons, some versions of Internet Explorer require the user to OK displaying "active content" when applets are launched. Just click when prompted to show content.

Java: Plug-in Java™ version 1.4.2_04 or later. A free update is available at `www.java.com`. If installing for the first time, choose "Typical Installation." With the Plug-in installed, make a (one-time only) cache setting. This is done as follows: from the desktop, double-click on the Java Plug-in icon (often in the Control Panel), and choose the Cache tab. In the window that appears, make sure the "Enable Caching" box is checked, then select the "Unlimited" button, and finally click on "Apply." (Your computer will probably use under 75 MB.)

Monitor: Super VGA Color Monitor with 256 colors and 800 by 600 or higher resolution.

Mac

Speed: 1 GH or faster (a G4 or G5 machine).

Operating System: Mac OS X 10.3.4 or later.

Memory: At least 128 MB RAM; 256 or more is recommended.

Browser: Safari 1.2 or later. Internet Explorer for Mac is *not* recommended for running these applets.

Java: Plug-in Java™ version 1.4.2_04 or later. A free update is available at `www.java.com`. Make a (one-time only) cache setting, which is done as follows: from the main startup screen, do: Go → Applications → Utilities → Java → Java 1.4.2 Plug-in Settings → Cache. Make sure the "Enable Caching" box is checked, then select the "Unlimited" button, and finally click on "Apply." OS X is smart, and will use only what's needed to make the applets run properly. This is usually under 75 MB.

Monitor: Super VGA Color Monitor with 256 colors and 800 by 600 or higher resolution.

To Run the Applets

There are two groups of applets—a set of 26 illustrating concepts from the book's Chapters; and a set of 10 for Appendix 1 (the "Formulas" appendix).

If either of the browsers recommended above is your default browser, then from the CD's "APPLETS" icon (which you'll see after double-clicking My Computer), launch by double-clicking on ChapterApplets.htm or AppendixApplets.htm. Otherwise, start the recommended browser and open ChapterApplets.htm or AppendixApplets.htm from File → Open...

In general, the speed at which the applets run increases somewhat if you copy the CD onto your hard drive, and run the applets from there.

Important Note: These applets are fairly memory intensive. For machines with 256 MB RAM or less, it is strongly recommended that you *quit the browser and restart it before running either Chapter or Appendix applets.* (You can quit Safari on a Mac by pressing ⌘ + Q.)

Best viewing: There are two settings that can markedly affect the appearance of browser-dependent programs: (i) Font size and (ii) Monitor resolution. These applets were designed to look good at the default settings of Medium font size and monitor resolution of either 800 by 600 or 1024 by 768. Your computer may be set differently; here's how to adjust these settings in the recommended browsers:

To adjust type size

PC: In the top menu bar of Internet Explorer, go View → Text Size.

Mac: In the top menu bar of Safari, click on the "A A - icon."

To adjust monitor resolution

PC: My Computer → Control Panel → Display → Settings (tab) → (Choose resolution here) → OK.

Mac: Click the little "Monitor icon" at the right of the top main menu bar and adjust resolution.

If this icon doesn't show, do this:

Go → Applications → System Preferences → Displays, and check the box "Show displays in menu bar" to activate this convenience.

Problems?

If the applets don't run, or seem to experience problems

- As noted above, some versions of Internet Explorer require the user to OK displaying "active content" when applets are launched. If you see such a message, simply click when prompted to show content.
- Make sure that the Java Plug-in is up to date. There is a big difference between v. 1.4.1 and v. 1.4.2, for example.
- It is especially important that the Java Plug-in cache size be set large enough. 75 MB will probably do, but "unlimited" is the safest.
- Check that the operating system is a recommended one. On a Mac, there are important differences between OS X 10.3.2 and OS X 10.3.3.
- Quit other applications—they consume memory, and can be an issue with certain computers. (A virus protection program running in the background is fine, however.)
- "Clean the slate" of the browser's memory cache—that is, quit and restart the browser before running either group of applets.

A Final Word

There are three varieties of applets

- **Maple animations** (Applets 9–10, 12–15). These are animated-gif applets—basically, a sequence of pictures quickly displayed one after the other to create the illusion of motion. These applets are not interactive.
- **JavaView applets** (Applets 1, 2, 11, 19, 22). JavaView is a powerful 3-D viewer, and can be used as an add-on library that interfaces nicely with Maple or Mathematica, for example. It greatly enhances the standard 3-D viewing capabilities, permitting three degrees of rotation, not just two, as well as uniform scaling and translation. It also allows great control over drawing elements—edge thickness and color being just two examples. To gain an appreciation of the wide range of capabilities of this important contribution to 3-D viewing, search for JavaView and explore the Home page as well as other sites.
- **Cabri applets** (All remaining applets). These are Java versions of programs written in the geometry-drawing program Cabri II. Those who own Cabri II or later can run the original programs after launching Cabri: from the File menu, navigate as follows: "Chapter" or "Appendix" folder → Cabri PAs → "Cabri Applet Items" folder → *.fig. You can animate (or multiple animate) any of the input points in Cabri figures. Due to a limitation in the Java 1.4.2 Plug-in, input points in the Appendix

applets usually restrict the figure to an ellipse. In Cabri itself, any such restricted input-slider's maximum value may be freed up, and you can then see the whole range of action. (From the Show/Hide menu item, Show the line segment's endpoint; then from Fix/Free, click to remove the tack. You can then drag to make the line segment longer.)

Index

Adams, John Q., 103
Affine
  equivalence, 16
  space, 15
  transformation, 16
Apollonius, 2, 309, 393
Asymptote, 22

Ball model, 15
Bezout, 186
Boundary point, 383

Camera obscura, 12
Cauchy, Augustin-Louis, 381, 383
Chandra X-ray telescope, 357
Chasles, Michel, 11
Circle of best fit, 240, 251
Confocal conics, 136
Continuous, 382
Copernicus, Nicolaus, 310
Curvature of a curve, 236, 242
Curvature of a surface, 245, 258

Dandelin, Germinal-Pierre, 86, 141
Da Vinci, Leonardo, 11
Dehomogenization, 35
Desargues, Gerard, 11
Descartes, René, 41, 306
Diffraction ruling, 348
Dirac, Paul Adrien Maurice, 75
Directrix, 107, 117, 122, 138, 364
Discriminant, 150, 161
Disk model, 15, 61, 194
Distance function, 105, 291, 294

Eccentricity, 85, 107, 126, 130, 135, 313, 361, 364
Eigenline, 153, 166
Eigenvector, 153
Einstein, Albert, 42, 68, 91, 111, 259–261, 263, 328
Euclid, 1
Euler, Leonhard, 186, 216, 217, 252, 380, 383
Eulerian angles, 259
Euler's theorem of 1760, 252

Fast Fourier Transforms, 356
Feynman, Richard, 283, 393
Foci, 93, 98
Focus, 4, 89, 104, 107, 364
Fundamental form
  first, 257
  second, 257
Fundamental theorem of algebra, 152, 200

Galileo Galilei, 311, 329, 337
Gauss curvature, 245, 248, 275
Gauss map, 245
General theory of relativity, 42, 68, 91, 260, 262, 263, 328
Genus formula, 232

Halley, Sir Edmond, 317, 321
Harish-Chandra, 164
Hausdorff, Felix, 382, 383
Homogeneous, 16, 17, 31, 33, 35, 60, 179
Huygens' principle, 345, 346
Huygens, Christiaan, 306, 345

Incidence/reflection property, 78, 79, 83, 103, 301
Interference fringes, 344, 346, 348, 351, 357

Jansky, Karl, 354

Kepler, Johannes, 4, 309, 312, 317, 337
Kepler's laws, 312, 313, 316, 332, 337, 338, 393

Latus rectum, 84
Leibnitz, G., 7
Lexicographic ordering, 76
Light cone, 67, 110
Line at infinity, 15, 16, 23, 29, 30, 122, 179, 185, 198, 202, 230
Linking number, 193, 213

Metric, 24, 27, 30, 51, 263, 265, 284
Meusnier's theorem, 272, 282
Michelson, Albert, 348, 352
Minkowski plane, 67, 68, 96, 108, 131, 162, 197, 218
Minkowski, Hermann, 68
Mormon tabernacle, 104
Multiplicity, 70, 152, 170, 175, 185, 186, 394

Newton, Sir Isaac, 42, 240, 263, 306, 316, 317, 321, 324, 326, 328, 337, 363
Null line, 100, 101, 109, 116, 197, 231
Nutation, 20

Open set, 17, 178, 179, 182, 381, 382
Order of a polynomial, 191, 201, 212
Orientable, 54, 179, 182

Pappus of Alexandria, 211
Pappus' theorem, 212, 213
Pascal, Blaise, 187
Pascal's theorem, 187, 203
Pinhole box, 12
Plato, 309, 310
Poincaré, Henri, 383
Point(s) at infinity, 10, 12, 14, 15, 20, 30, 31, 46, 62, 132, 175, 176, 178, 179, 194
Polar equation, 314, 315, 331, 360, 363, 378
Poncelet, Jean-Victor, 22
Precession, 327, 394
Principia, 317, 321
Projective
  geometry, 22, 178, 209, 379
  line, 30, 178
  plane, 9–11, 29
  space, 14, 15, 178
  transformation, 16, 20, 178

Quadratic form, 150
Quetelet, Adolphe, 86, 141

Radio telescope, 355
Reber, Grote, 354
Retrograde motion, 310–312
Riemann, G. F. Bernhard, 383
Riemann sphere, 133, 135, 233

Schroeder, Erwin, 75
Semi-latus-rectum, 84, 314, 361, 364, 370, 375, 378
Separation (squared), 66, 67, 96, 109, 130, 197, 218
Sphere model, 11, 14, 15, 58
Spherical Pythagorean theorem, 286
Square-root construction, 308
Stakes and string method, 4, 5, 86, 103, 287, 339, 341, 366, 376
Statuary Hall, 103
Stereographic projection, 133

Theorema egregium, 264–6, 282
Three tangents theorem, 213
Topological manifold, 386
Topological space, 382
Tractrix, 122, 266
Tycho Brahe, 312, 337

Very Large Array (VLA), 356, 357

Weyl, Hermann, 90
Whitney, Hassler, xii, 132
Wren, Christopher, 317

About the Author

Keith Kendig graduated *Summa cum Laude* and Phi Beta Kappa from UCLA, where he continued with graduate work, earning his Ph.D. under Basil Gordon. He subsequently spent two years at the Institute for Advanced Study in Princeton, working with Hassler Whitney.

Keith is the author of *Elementary Algebraic Geometry*, which appeared as Volume 44 in Springer-Verlag's Graduate Texts in Mathematics Series. His article "Is a 2000-Year-Old Formula Still Keeping Some Secrets?" won the MAA's Lester Ford award in 2000. Although his field is algebraic geometry, he frequently teaches graduate courses in applied mathematics. He's a firm believer in using computers in both teaching and research.

Keith is a vegetarian, and co-authored a book on vegetarian cooking with his wife Joan. He's a "jogger for all seasons," as well as a semi-professional cellist.